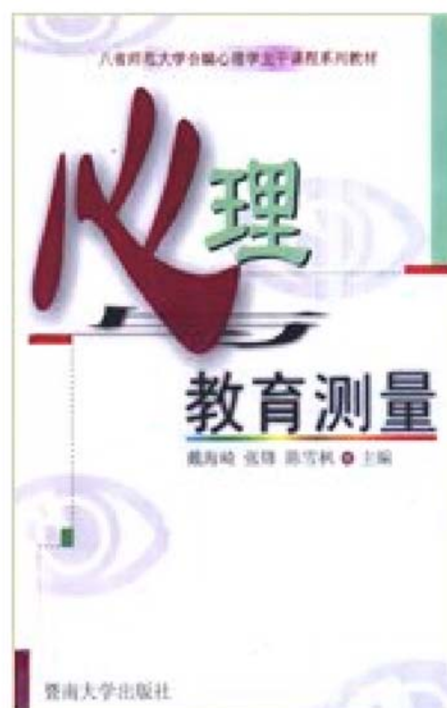

八省师范大学合编心理学主干课程系列教材
心理与教育测量
戴海崎 张锋 陈雪枫 主编
暨南大学出版社

心理学主干课程系列教材

编委会成员名单

·八省师范大学合编心理学主干课程系列教材·

心理与教育测量

主编 戴海崎 张锋 陈雪枫

华南师大心理系教材建设基金资助出版

暨南大学出版社

图书在版编目（CIP）数据

心理与教育测量/戴海崎，张锋，陈雪枫主编．—广州：暨南大学出版社，1999.12
ISBN7-81029-696-5

Ⅰ．心…
Ⅱ．①戴…②张…③陈…
Ⅲ．教育心理学
Ⅳ．G449.1

书名：心理与教育测量
责编：陈 红
出版：暨南大学出版社出版（广州·石牌 510632）
排版：暨南大学出版社照排中心
印刷：东莞市印刷厂印刷
发行：新华书店
开本：850×1168 1/32
印张：14.875
字数：37万
版次：1999年2月第1版
印次：1999年12月第2次印刷
印数：3001-8000册

定价：24.00元

编写说明

《心理与教育测量》是华南师范大学心理学系组织南方8省师范大学编写的心理学主干课程系列教材之一。本书从测量学基本原理、测验编制技术、知名测验性能3个方面总结前人所编教材的经验，力求反映测量研究领域的当代特色。本书主要有如下特色：

1. 辟专章介绍目标参照测验的理论与技术。

2. 介绍了认知心理学的一些测量学新观点。

3. 增加了测验等值、题库建设、教师自编测验等实用技术的介绍。

4. 加强了对我国学者在测量学领域研究活动与成果的介绍。

5. 专章介绍了现代测量理论两个主要分支项目：反应理论与概化理论的新发展。

本书可作为心理学、教育学、社会学等专业的测量课教材，也作为从事心理咨询、考试评价、人员测评等工作的人员的参考书。

全书体系是在主编提供的框架基础上集体讨论而定的，编写分工如下：

张　锋：第一、八、十四章

罗黎辉：第二章

杨志明：第三、四、五章

龙文祥：第六、七章

戴海崎：第九、十一、十七章

陈雪枫：第十、十二、十三章

龙立荣：第十五、十六章

本书初稿由戴海崎、张锋分工审阅提出修改意见，最后由戴海崎统校定稿。

本书的编写得到心理学主干课程教材编委会的指导，特别是受到了编委会主任莫雷教授的格外关心；江西师大、云南师大、华中师大、湖南师大、安徽师大等校均对本书的编写出版给了很大帮助，在此一并深表谢意。我们还得感谢暨南大学出版社对本书的扶持。在编写中，我们参考了国内外大量资料，有些还作了摘引，在此也向这些作者表示感谢。

编　者

1997.7.1.

目　录

第一章　心理与教育测量概论

本章提要：

●测量的基本性质及其要素
●测量量表的四种水平
●心理与教育测量及其理论基础
●心理与教育测量的量表和测验
●心理与教育测量科学研究和实际工作中的意义
●心理与教育测量工作者的素质要求与道德准则

心理与教育测量是我国各大学心理专业和教育专业学生必修的重要的专业课，它在心理科学、教育科学的基础学科和应用学科之间起着一种中介作用。一方面，它是基础心理学科和基础教育学科的深化，是从事基础理论研究的方法课；另一方面，它又是应用心理学科和应用教育学科的基础，是从事实际应用研究的工具课。在本章里，我们将讨论心理与教育测量的若干基本概念和基本问题，以便为学习以后各章的具体知识提供一个基本的框架。

第一节　一般测量概述

一、测量及其种类

测量（measurement）是人类生产和生活中普遍存在的现象。农业生产要丈量土地面积，工业生产要测定产品的技术指标，地质勘探要测定海拔高度和地质指标，医疗工作要测定人体的生理指标，教育工作要测定学生的学业成绩。至于科学研究中的测量活动就更加普遍，也更加严格了。那么，究竟什么是测量呢?

简单地说，测量就是依据一定的法则使用量具对事物的特征进行定量描述的过程。

所谓“一定的法则”，是指任何测量都要建立在科学规则和科学原理基础之上，并通过科学的方法和程序完成测量过程。例如，用杆秤测量物体的重量，依据的是物理学上的杠杆

原理；用温度计测量温度，依据的是热胀冷缩原理。有的测量依据的法则比较科学和完善，测量的结果比较准确、可靠，而有的测量依据的法则比较粗糙和欠成熟，测量结果的准确性和可靠性较差。有的测量依据的法则的操作比较直观和简单，一般的人不需要经过专业训练就很容易掌握，而有的测量所依据的法则的操作复杂，需要经过专门训练才能逐步掌握。

所谓“事物的特征”，是指所要测量的事物的特定属性。例如，物体的重量、长短、高矮；物体运动的速度；物体中某些特定成分的含量等等。这些不同的特征就是测量的特定对象。一种事物有各种各样的特征，对不同的特征要用不同的测量工具、依据不同的法则进行测量。有些事物的特征直观明显，具有外显性（如物体的重量、长度等），所以在测量中容易被确定，测量的结果具有无可争辩性，容易被所有的人认同和接受；而有些事物的特征不那么外露，具有内隐性（如人的智力水平、性格特点等），所以在测量中难以准确界定，测量的结果不容易获得清楚的解释，因而也不容易取得多数人的认同和接受。

所谓“量具”，是指测量中所使用的工具。例如，重量测量中的杠秤、电子秤，长度测量中的木尺、皮尺，体温测量中的体温计等等。不同的测量要用不同的量具，不同量具所使用的单位和参照点也不同。

所谓“定量描述”，是指任何测量的结果总是对事物特征的量的确定。虽然有时人们把诸如“1”代表男，“0”代表女这样的做法也叫做测量，但这里的数字仅仅是一种符号，并不是数量。所谓“数量”不仅指事物特征的符号，而且指一种有序的量。数量具有4个特征：一是它的区分性，即一个数（如“1”）不同于另一个数（如“2”）；二是它的序列性，即 $1<2<3<4\cdots\cdots$；三是它的等距性，即 $2-1=1$，$3-2=1$，所以，

2 - 1 = 3 - 2；四是它的可加性，即一个数加另一个数产生第三个数。数的这些特点是一切数学运算的基础，同样，也正是数的这些特点使得对事物特征的差异的测量成为可能。有的测量对事物的特征定量描述的精确度高些，而另一些测量对事物的特征定量描述的精确度差些。测量的精确度既与测量对象的性质有关，也同测量时所用的工具有关。

首先，测量的精确度决定于测量对象本身的性质。我们可根据测量对象的性质把它分为 3 种类型：①确定型，即在一定条件下，事物的量保持恒定不变。例如，物体的长度和重量，只要物质的温度不变，受力状况不变，其长度也就不会改变；只要物体在地球表面的水平位置和垂直高度不变，其重量也不会改变。②随机型，即事物的量随机改变。例如，人的短时记忆的容量，尽管实验者在实验过程中每次向被试呈现刺激的条件保持恒定，但每次测量的结果总是存在差异，不过，这种差异又总是保持在一定的范围内，量的改变趋势也呈现出一定的规律。③模糊型，即事物的量本身就是模糊不定的，难以获得确定的量。例如，对人的性格特征，尽管人们习惯于用热情奔放或冷若冰霜等词汇来描绘，而且也能够区分出两个同是热情奔放的人在程度上的差别，但这种差别的量却是很模糊的。显然，对确定型的事物进行定量描述要比对随机型和模糊型的事物进行定量描述要容易得多，因此，测量的精确度也要高得多。但是，即使对确定型的事物也不能做出绝对精确的描述，在任何测量过程中都会有误差存在，所不同的是误差的大小而已。

其次，测量的精确度决定于测量工具（量尺）的精密性。不言而喻，使用技术上完善的测量工具比使用技术上粗糙的测量工具，其测量结果要精确得多。对于长度的测量，用皮尺测量比之用脚步测量，其结果要精确得多；而用激光测量比之用

皮尺测量，其结果又要精确得多。同样，对于重量的测量，用杆秤测量比个人的主观估计，其结果要精确得多；而用电子秤测量比用杆秤测量，其结果又要精确得多。因此，尽可能使用技术精密的测量工具，是保证测量精确度的重要条件。但是，不论使用何等精密的测量工具，实际测量中仍然会有误差存在，所不同的也仅是误差的大小而已。测量学的目标之一是设法尽可能把误差减少到最低程度，而不可能完全消灭误差。

测量技术被广泛用于工农业生产、商业活动、科学研究和人们的日常生活领域。根据测量对象的性质和特点，可以将各种不同形式的测量大致分为4种类型：①物理测量：即对事物的物理特征的测量。如长度测量、重量测量、面积测量、速度测量等等均属物理测量。②生理测量：即对机体生理特征的测量。如对动植物各种化学成分含量的测量，对人体各种生理机能的测量等等均属生理测量。③社会测量：即对社会现象的测量。如在人口普查、经济统计、民意调查中所使用的测量技术均属社会测量。④心理测量：即对人的心理特征的测量。如智力测量、人格测量、职业兴趣测量、态度测量等等均属心理测量。狭义的教育测量主要指对学生学业成绩和知识水平的测量，此时，教育测量可以被包括在心理测量的范畴之内。但是，广义上的教育测量不仅包括对学生学业成绩和知识水平的测量，而且包括对教育领域中其他教育现象的测量。如对教师教学水平的测量、对整个学校办学质量的测量、对学校管理水平的测量等等。此时，教育测量当属社会测量的范畴。本书所使用的教育测量是指狭义的教育测量，但为了与其他心理测量有所区别，将教育测量与心理测量这两个术语并列使用。

二、测量的基本要素

任何测量都必须具备两个基本要素，即测量的参照点和测量的单位。

(一) 测量的参照点

从根本上说，测量是确定特定事物的特定特征的数量。因此，在测量工作中，必须有一个量的起点。这个起点就叫作测量的参照点。要使两个测量结果能够相互比较，必须使这两个测量建立在同一个参照点上。因为参照点不同的两个测量，其结果的意义完全不同，没有可比较的共同基础。

参照点有两种：一种是绝对参照点，即以绝对的零点作为测量的起点。如长度测量和重量测量就是建立在以绝对的零点为参照点的基础上的测量。这个绝对的零点的意义就是“无”，即没有重量或没有长度。以此为测量的起点，去确定某种事物有多重或有多长。另一种是相对参照点，即以人为确定的零点为测量的起点。如对地势高度的测量，就是以海平面为测量的起点。此时，人们假定海平面的高度为“零”，然后去确定陆地高出海平面多少，再如对气温的测量，是以水的冰点为测量的起点。此时，人们假定水刚刚能够结为冰的温度为“零”，然后确定气温高于或低于“零”多少度。

最为理想的测量参照点当然是绝对参照点，因为它的意义最为明确。但在许多情况下，人们难以找到绝对参照点，所以必须改用相对参照点。采用相对参照点为测量起点的测量结果只能进行加减运算，而不能进行乘除运算　因此，它的两个值

之间没有倍数关系。例如，在智力测量中，假定甲的智商为100，而乙的智商为50，我们不能说甲的智商是乙的智商的2倍，而只能说甲的智商高出乙的智商50。

（二）测量的单位

测量的第二个基本要素是它的单位。不同测量所用的单位是不同的。长度测量的单位是毫米、厘米、分米、米等等，而重量测量的单位是毫克、克、千克、吨等等。理想的测量单位应当具备两个条件：一是要有确定的意义，即对同一单位，所有的人的理解都是相同的，不允许作出不同的解释。例如，所有的人对重量单位“千克”的解释都是一样的，没有歧义。二是要有相等的价值，即第一个单位与第二个单位之间的距离等于第二个单位与第三个单位之间的距离。例如，30公斤与20公斤之差等于40公斤与30公斤之差。但是，在某些情况下，要具备这两个条件是相当困难的。例如，教育与心理测量中的单位就往往难以达到这个要求，它远没有其他测量中使用的单位成熟和完善。这一点我们在后面还会谈到。

三、测量的量表

要测量某一特定事物的特定特征的数量，必须首先选择一个具有确定单位和测量参照点的数字连续体，将欲测量的特征与这个连续体相比照，确定它的位置，看它距参照点的远近，就会得到该特征的一个测量值。这种能够使事物的特征数量化的数字的连续体就是量表（scale），制定量表的单位和参照点不同，就会编制出不同的量表；不同的量表具有不同的测量水

平，相应地测量的精确度也不同。根据测量的不同水平以及测量中使用的不同单位和参照点，我们把测量量表分为 4 种。

（一）命名量表

命名量表（nominal scale）是最低水平的测量量表，它只是用数字代表事物或用数字对事物进行分类。在这种情况下，数字只是事物的符号，而没有任何数量的意义。因此，运用命名量表时不能作常用的数量化分析。如我们不能说 6 号学生 > 5 号学生 > 4 号学生，当然也不能进行代数运算。正因为如此，有人认为运用命名量表进行的测量不能算作是真正意义上的测量。

命名量表又可细分为两种形式：一是名称量表，即用数字指代个别事物。如用数字给学生或运动员编号。二是类别量表，即用数字指代事物的种类。如用 1、2、3、4、5……分别代表不同的职业。

适合于对命名量表进行统计分析的统计方法有百分比、次数、众数和 χ^2 检验。

（二）顺序量表

顺序量表（ordinal scale）是次低水平的测量量表，它不仅能够指代事物类别，而且能够表明不同类别的大小、等级或事物具有某种特征的程度。各种比赛、评估中的名次排列就是一种典型的运用顺序量表进行的测量。例如，在各种体育比赛中，我们通常取前 3 名，分别用 1、2、3 代表，那么，我们就可以说，1 > 2 > 3。这表示，第 1 名的水平高于第 2 名的水平，第 2 名的水平又高于第 3 名的水平。这种按照事物的大小、等级、程度而排列数字的量表就叫顺序量表。

在顺序量表中，数字只表示等级、大小和程度的顺序，它

既没有相等的单位，也没有绝对的零点。换言之，它既不表示事物特征的真正的数量，也不表示绝对的数值，因此不能进行代数运算。

适合于对顺序量表进行统计分析的统计方法有中位数、百分位数、等级相关系数和肯德尔和谐系数等。

（三）等距量表

等距量表（equal interval scale）是较高水平的测量量表，因为它不仅能够指代事物的类别、等级，而且具有相等的单位。等距量表的数字是一个真正的数量，这个数量中各个部分的单位是相等的，因此就可以对其进行加减运算。例如，在测定气温时，10℃和15℃的差别与15℃和20℃的差别是相等的。

等距量表没有绝对的零点，它的零点是人们假定的相对零点。因此，对于等距量表中的两个数量不能进行乘除运算，它们之间不存在倍数关系。例如，我们不能说20℃是10℃的两倍。

适合于对等距量表进行统计分析的统计方法有平均数、标准差、积差相关系数以及t检验和f检验。

（四）比率量表

比率量表（ratio scale）是最高水平的测量量表，因为它除了具有类别、等级、等距的特征外，还具有绝对的零点。使用比率量表，不仅可以知道测量对象之间相差的程度，而且可以知道它们之间的比例。在长度测量和重量测量等物理测量中，人们广泛使用比率量表。例如，在长度测量中，测得甲的长度为9米，乙的长度为3米。此时我们不仅了解到甲比乙长6米，也了解到甲的长度是乙的长度的3倍。在重量测量中，测得甲的重量为40公斤，乙的重量为20公斤。那么，我们既可知道甲比乙重20公斤，又可知道甲的重量是乙的重量的两倍

适合于对顺序量表进行统计分析的统计方法除了与等距量表相同外，还与几何平均数、变异系数相同。

第二节　心理与教育测量的性质

一、心理与教育测量的定义

根据一般测量的定义，我们可以将心理与教育测量定义为：依据一定的心理学和教育学理论，使用测验对人的心理特质和教育成就进行定量描述的过程。与一般测量的定义相比较，心理与教育测量的定义既具有一般测量的共同属性，又具有其独特的性质。

首先，心理与教育测量依据的法则在很大程度上只是一种理论，很难达到如同物理测量依据的法则那样普遍被人们接受的水平。心理与教育测量学家凭借这些理论来编制测量的工具并完成测量工作。例如，测量学家在编制智力测验时对智力本身的性质存有非常不同的看法，对智力的结构成分也有非常不同的理解。因此，从一种智力测验上得到的测量分数与另一种智力测验上得到的测量分数可能具有不同的意义。由于所依据的法则不够成熟，即使使用同一种测验测量，所得结果也不像物理测量那样准确和可靠。这种情况在人格测量领域表现得更加明显。

其次，心理与教育测量的对象是人的心理特质和教育成就。教育成就的含义比较明显，是指一个人通过接受教育而获

得的知识水平。心理特质的含义则比较含混，不大能够给予清楚的界定。心理学上通常将“特质”理解为相对稳定的、对个人的行为具有持久的调节作用的心理特征，如智力、兴趣、态度、人格等等均可以视为特质。智力、兴趣、态度、人格等特质本身就是很抽象的概念，在测量工作中，将其具体为可操作的测量对象本身就是一件非常复杂的工作。心理特质显然具有内隐性，我们不可能像测量重量或长度那样直接测量人的心理特质的量，而是通过测量个人在特定情境中的外显行为来推断他的心理特质。这就决定了心理与教育测量只能是一种间接测量。

再次，心理与教育测量的量具是由有关领域的专家编制，经过长期的试用、修订、完善而逐渐形成的标准化测验(test)。它的编制是一项高度专门化的系统工作，要达到科学所要求的水平绝非易事。

最后，心理与教育测量的目标虽然是对人的心理特质和教育成就进行定量分析，但这种定量分析的精确度远不及物理测量的精确度高。这首先是由人的心理特质的高度复杂性所决定的，同时也与目前的测验编制理论不够先进，测验编制的技术水平不高有关。

二、心理与教育测量的理论基础

在阐述心理与教育测量的定义的过程中，我们在说明心理与教育测量特征的同时，着重强调了心理与教育测量的复杂性和难度。也许正因为如此，人们对心理与教育测量是否必要和是否可能持有怀疑态度。归纳起来，怀疑心理与教育测量的必

要性和可能性的理由不外乎两个方面：①人的心理现象和知识水平是一种主观存在，它的复杂性、流动性和内隐性的特征使人们不可能对其进行直接测量。②目前的心理与教育测量的技术手段远未达到如物理测量那样的准确和可靠。

那么，心理与教育测量究竟是否必要和是否可能？对此，我国古代学者孟子早在2000多年前就给予了明确的、并且是肯定的回答（参见第二章）。但直到本世纪初期，随着心理与教育测量运动的发展，这个问题才真正摆在了测量学家的面前。1918年，桑代克曾提出，“凡客观存在的事物都有其数量”。1939年，麦柯尔进一步指出，“凡有其数量的事物都可以测量”这两个命题被公认为是心理与教育测量的理论基础。

从辩证唯物主义的观点看，任何事物都是质和量的统一，事物的质的差异是分类的前提，而事物的量的差异则是测量的前提。这里的“事物”不仅指外在的客观现象，而且指人的内在的主观现象。根据这一前提，我们认为：①人的心理现象和知识水平如同其他一切物理现象一样是有差异的，这种差异不仅包含质的方面，也包含量的方面。因为有差异，所以有必要测定差异的数量，描述差异的程度。②心理特质和知识水平虽然不是物理实体，不能直接测量，却必然要表现于人的外部行为之中，并调节着人的外部行为。因此，通过观测人的外部行为的差异就有可能测量出人的心理特质和知识水平的差异。③心理与教育测量的准确性、可靠性和精确度如同其他一切测量技术一样是相对的，也同其他一切测量技术一样必然随着科学技术的进步和发展而逐步提高。目前的心理与教育测量的科学性还达不到人们所期望的准确的高度，但测量不准不等于不能测量，目前测量不准也不意味着将来永远测量不准。近100年来，心理与教育测量学家正是抱着这种信念进行了大量的研究，取得了明显的成效，已经初步形成了一套比较科学的测量

理论与技术。可以设想，随着研究工作的拓展和深化，心理与教育测量的科学水平将会进一步提高。

三、心理与教育测量的量表

我们在第一节里已经谈到，测量中所使用的单位和参照点的水平不同，就会有不同水平的测量量表。那么，心理与教育测量的量表属于哪一水平呢？从本质上讲，心理与教育测量的量表属于顺序量表。这是因为：①从所使用的参照点来说，教育测量和心理测量领域的参照点均为相对参照点。例如，在学期末的学科考试中，通常的做法是把学生的成绩确定在0～100分之间。显然，这个0分是人为假定的起点。因为即使某学生得了0分，我们也不能说该生在本学期内没有学到任何知识，或者说该生的知识水平为“零”。在智力测量中，假定某一儿童不会做任何一个题目，那么，他的成绩为0分，但这个0分也并不表示他的智力水平为“零”。这就决定了心理与教育测量的量表不可能达到比率量表的水平。②从所使用的单位来说，教育与心理测量的单位远没有其他测量的单位成熟和完善。一是教育与心理测量所使用的单位的意义不太明确。例如，在各种形式的考试中，虽然使用单位都是“分”，但实际上，数学考试中的“分”和语文考试中的“分”的意义是不相同的。学生在不同学科上的考试成绩所代表的不是同一个东西。二是在教育与心理测量中的单位常常不等值。例如，同一次数学考试，学生做对一道较简单的题目，得到1分，同样做对一道较复杂的题目，也得到1分。从表面上看，前者的1分和后者的1分是等值的。但实际上，它们所反映的学生的知识

水平是不相等的。由于单位的意义不同，单位的价值不相等，所以各科的考试成绩不能直接相加而求出总分，也不能根据总分求各科平均分。这就决定了心理与教育测量的量表不是直接的等距量表。

由于顺序量表的参照点没有绝对零点，而且它的单位不等值，大量的统计方法不能直接应用到顺序量表的分数上去，因此在理论研究和实际应用工作中受到极大的限制。为了克服这些缺陷，心理与教育测量学家希望将顺序量表上得到的分数转化到等距量表上去解释。也就是说，希望采用统计方法把顺序量表的分数转换到具有相等单位的等距量表上。为此，教育与心理测量学家做了大量的研究工作，收到了一定的成效。目前，大多数心理与教育测量的分数解释工作是在等距量表上进行的，但是，很难说这项工作在本质上改变了教育与心理测量分数单位的不等值性。

四、心理与教育测量中的测验

如前所述，心理与教育测量工作是在测验上完成的，而测验是由有关领域的专家经过长期的编制、试用、修订、完善而逐渐形成的标准化测量工具。对于什么是测验的问题，学术界尚未取得一致的意见，被多数测量学家所接受的定义是美国心理测量学家阿娜斯塔西（A·Anastasi）提出的定义。她认为，“心理测验实质上是对行为样本的客观的和标准化的测量。”根据这一定义，编制一个测验应当具备下列4个基本条件。

(一) 行为样本

抽样是测量学上普遍采用的方法。例如，在水质检验中，检验人员在要检验的水中抽取一小部分样水予以测定，根据对样水的测定结果推断整个水的质量；在医疗验血中，医生也只是抽取很少一部分血样进行测定，并根据此测定结果推断其整体的情况。从整体中抽取出来作为测量对象的一组样品叫作样本。如上所说，心理与教育测量是间接测量，是通过测量人的外部行为来推断人的心理特质和教育成就。但人的行为是多种多样的，要把人的所有行为都作为测量的对象显然是不可能的，也是不必要的。为此，测量学家的做法是从人的大量行为中抽取与欲测量的心理特质直接有关的一组行为进行测量，并依据对这一组行为的测量结果推断其心理特质和教育成就。这一组被抽取出来的、作为直接的测量对象的行为即是行为样本(sample of behavior)。例如，我们要知道学生的数学运算能力的高低，就可能选择若干有代表性的数学问题，要求学生解答这些问题。学生在解答这些数学问题时的行为就是我们要测量的直接对象，当我们根据这一组行为来推断其整体的数学运算能力时，这一组行为就是数学运算能力的行为样本，而引起学生行为的那些数学运算问题就是测验。所以，简单地说，测验就是引起特定行为的工具。显然，所抽取的行为样本必须是能够给测量人员提供有意义的、足以反映个人特定心理特质的一组行为，而要做到这一点，首先要使构成测验的项目与要测量的行为有关。

(二) 标准化

标准化（standardization）是编制测验的一个重要步骤，也是测验的重要条件。为了使接受测量的不同个人所获得的分

数有比较的可能性，测验的条件必须对所有的个人都是相同的。在相同的测验情境中，唯一的自变量是正在接受测量的个人的心理特质，这样，测量结果才具有客观性。测验的标准化就是指测验的编制、实施、记分以及测量分数的解释的程序的一致性。测验的标准化需要具备下列条件：

1．测验内容的标准化

标准化的首要前提，是所有接受测量的个人实施相同的或等值的测验内容。测验内容不同，所测得的结果便没有可比较的基础。

2．施测条件的标准化

标准化的第二个条件，是所有接受测量的个人必须在相同的施测条件下接受测验。其中包括：①相同的测验情境。②相同的指导语。③相同的测验时限。

3．评分规则的标准化

评分规则的标准化要求评分结果具有客观性，只有当评分的结果具备了客观性，才能将测量分数的差异归之于个人心理特质和知识水平的差异。为此，测验中所制定的评分规则要足以使不同的评分人的评分结果保持最大程度的一致。

4．测验常模的标准化

编制测验的一个重要步骤是编制测验的常模（norm）。在心理与教育测量领域，由于测量分数没有绝对的零点作为参照点，所以，孤立地看待一个测量分数是没有什么意义的，必须将该测量分数与他人的测量分数相比较，才显示出它的意义。常模的功能就是给解释测量分数提供一个可比较的参照点。在许多情况下，常模是一组有代表性的被试群体的平均测验分数。这个平均测验分数表示的是普通人的一般状况。解释个人的测量分数就是将这一分数与常模分数相比较，看该分数高于或低于常模分数多少。例如，在能力测量领域，如果某一个人

的测量分数高于常模分数，则此人的能力水平高于普通人的平均水平；相反，如果某一个人的测量分数低于常模分数，则此人的能力水平低于普通人的平均水平。

常模既然是一组有代表性的被试群体的平均测验分数，那么编制测验常模的关键是要抽取有代表性的被试样本，它要求按照科学的抽样原则抽取样本中的每一个个体。

这里需要特别说明的是，不要把常模的概念与我们通常理解的标准的概念混淆起来。标准指的是理想上期望达到的程度，而常模指的是被试群体实际达到的程度。以常模为参照编制的测验叫常模参照测验，以标准为参照编制的测验叫目标参照测验或标准参照测验。在此主要讨论常模参照测验，但也涉及到目标参照测验。

（三）难度或应答率

在编制教育成就测验和各种形式的能力测验时，一个很重要的指标是确定项目的难度值。测验项目是按照其难度值由简单到复杂编排的，而项目的难度是通过计算被试答对某一项目的人数比例来确定的。例如，比内—西蒙智力量表（1905）中的 30 道题目就是根据 50 个智力正常儿童和少数智力落后儿童接受该测验的结果而编排的。这是最早用客观方法决定项目难度的尝试。难度太低或太高都不能有效地将不同水平的个体区分开来，从而也就不能保证测验的科学性。

编制诸如态度测验、兴趣测验、性格测验不存在难度问题，却有一个对项目的应答率问题。如果在某些项目上，答“是”或答“否”的被试人数太多或太少，则同样不能有效地区分不同态度、兴趣或性格的人。而应答率也必须通过客观的统计计算确定。

(四)信度和效度

评价一个测验是否科学的重要指标是它的信度和效度。信度指的是一个测验的可靠性，即用同一测验多次测量同一团体所得结果之间的一致性程度。我们用钢片卷尺去测量一木杆的长度，所得结果是可靠的，因为无论是由一个人数次测量，还是分别由数个人去测量，所测得的结果都是高度一致的。如果改用橡皮软尺去测量木杆，一个人多次测量或多人测量的结果就难得高度一致。这就是说，橡皮软尺这种测量工具的信度不高。由此可见，信度是衡量测验科学性的最基本的指标。效度指的是一个测验的有效性，即一个测验在多大程度上能够测到它所要测量的心理特质。如果一个测验所测得的不是它所要测得的特质，则这个测验就是无效的。例如，智力测验所要测得的特质应该是智力，如果一个智力测验测到的不是智力，而是知识，那么无论它的信度有多高，这个智力测验对于测量智力都是无效的。由此可见，效度是衡量测验科学性的最重要的指标。

第三节　心理与教育测验的种类及其功能

一、心理与教育测验的种类

为了满足心理与教育测量工作的需要，近百年来，测量学家编制了大量的测验，涉及到各个方面和各个领域。这就有必

要对各种各样的测验进行分类。采用的分类标准不同，就会有不同的测验分类系统。

(一) 按测量对象所作的分类

1. 智力测验

旨在测量个人的智力（一般认知能力）水平的高低。这是心理测量最早涉及的领域，也是目前发展得相对成熟的一种测验。国内外比较著名的智力测验有“斯坦福—比内量表”、“韦克斯勒智力量表”、“瑞文推理测验”等等。

2. 能力倾向测验

旨在测量个人的潜在的才能，预测个人的能力发展倾向。能力倾向测验一般可分为两种：一种是一般能力倾向测验，测量个人多方面的潜能；另一种是特殊能力倾向测验，测量个人的特殊潜在能力，如音乐能力倾向测验、机械能力倾向测验等。

3. 成就测验

旨在测量个人在接受教育后的学业成就。成就测验有两种类型：一是学科成就测验，测量受教育者在某一科目上的学习成就；二是综合成就测验，测量受教育者在各学科上的综合学业成就。

4. 人格测验

旨在测量个人在诸如兴趣、态度、动机、气质、性格等方面的心理特征。由于人格一词的含义太广泛，一个具体的测验不可能含盖如此广泛的内容，所以常常有偏重，也有测量单一人格特质的测验，如内向——外向量表。人格测验又主要分为两类：一类是自陈人格问卷，比较著名的有“明尼苏达多相人格调查表”、“卡特尔 16PF 测验”、“艾森克人格问卷”等；另一类是投射测验，如“罗夏克墨迹测验”、“主题统觉测验”等

（二）按测量方式所作的分类

1. 个别测验

对于有些测验，同一主试在同一时间内只能测量一个被试，所以被称为个别测验。例如“斯坦福—比内量表”、“韦克斯勒智力量表”等智力测验以及“罗夏克墨迹测验”、“主题统觉测验”等人格测验均属个别测验。个别测验有许多优点，一是主试对被试的做题行为有仔细的观察，有机会获得测量分数之外的信息；二是主试与被试面对面交流的机会更多，容易与被试建立起融洽的合作关系；三是对于一些特殊被试（如幼儿、文盲），只能采用个别测量，以便主试代替被试记录其行为反应。但个别测验也有它的缺点，一是费时间，难以在短时间收集大量的测量资料；二是测验手续比较复杂，需经过较高水平训练的人担任主试。

2. 团体测验

对于有些测验，同一主试在同一时间内能够测量许多被试，所以被称为团体测验。例如，“瑞文推理测验”、“陆军甲、乙种团体智力测验”以及绝大多数自陈人格问卷均属于团体测验。团体测验的优点是节省时间，可以在短期内收集到大量的测量数据，所以在诸如教育、人事选拔、团体比较研究中被广泛使用。它的缺点是由于同一时间内接受测量的被试多，不易有效地控制被试的行为，容易产生测量误差，从而影响测量的信度和效度。

团体测验可用于个别测量，但个别测验不能用于团体测量。

（三）按测验内容的形式的分类

1. 文字（纸笔）测验

有些测验的内容是通过文字的形式表现的，被试也用文字作答，所以被称为文字测验，也叫纸笔测验。此种测验实施起来方便，团体测验多采用此种方式编制。其缺点是容易受被试的文化背景的影响，从而降低测验的效度。

2. 非文字（操作）测验

有些测验的内容是通过图形、仪器、工具、实物、模型等形式表现的，被试通过指认、手工操作向主试提供答案，所以被称为非文字测验或操作测验。此种测验不受或少受文化背景的影响，因此，在设计所谓“文化公平测验”时常采用这种方式。同时，也适用于测量学前儿童及文盲的心理特质。但非文字测验常局限于个别测量，在时间上不经济。

有些测验（如“斯坦福—比内量表”、“韦克斯勒智力量表”）既包括了文字测验的项目，也包括了操作测验的项目。

（四）按测验功能所作的分类

1. 成就测验与预测测验

成就测验的目的是测量个人在某一领域已经达到的实际成就。而预测测验的目的在于测量个人在未来某一方面获得成功的可能性。

2. 难度测验与速度测验

难度测验的功能在于识别个人能够达到的最高水平。通常包括各种难度不等的项目，其中有一些极难的项目，由易到难排列，供各种不同水平的被试作答。速度测验的功能在于识别个人做题的最快速度。通常包括大量相对容易的项目，要求被试在严格限定的时间内作答，被试在规定的时间内答对的题数

越多，则表示他的反应越快。

3. 描述测验与诊断测验

描述测验的功能在于通过测量来描述某一特定群体在某一心理特质上的一般状况。例如，心理学上关于智力发展趋势的研究、关于智商在不同年龄阶段的稳定性的研究、关于智力水平与学业成就关系的研究、关于男女智力差异的研究等都是通过运用智力测验而完成的。这些测验研究的目的都是为了描述和说明一个实际问题。诊断测验的功能是对个人的问题行为及其原因进行诊断。这种测验通常在教育和临床治疗领域被广泛应用。例如，学生学业成绩不良的原因可能是多种多样的，究竟是什么原因，需要运用诊断测验才能弄清楚。

(五) 按评价所参照的标准分类

1. 常模参照测验

常模参照测验将被试水平与常模相比较，以评价被试在团体中的相对地位为目的。

2. 目标参照测验

目标参照测验将被试水平与一绝对标准相比较，以评价被试有无达到该标准为目的，也称标准参照测验。

3. 潜力参照测验

潜力参照测验将被试水平与自身潜力相比较，以评价被试有无充分发挥自身潜力为目的。

二、心理与教育测验的功能

心理与教育测验经过将近100年的发展，现已被广泛地应

用于科学研究和教育、临床、人才选拔等实践领域，并发挥出日益重要的作用。概括地说，心理与教育测验的功能主要表现在两大方面。

（一）理论研究功能

1. 收集研究资料

在心理学和教育学的许多研究工作中，都需要通过测验来获得第一手资料。例如，为了查明影响学生学业成绩的心理因素，我们需要运用智力测验、学习能力倾向测验、成就动机测验、学习兴趣测验、人格测验和学业成就测验，通过计算各种心理因素的测量分数与学业成就测验的分类之间的相关系数进行回归分析，然后根据测验所获得的实证资料做出科学结论。

2. 建立和检验理论假设

在心理学的研究中，通常需要根据已有的测验研究成果提出理论假设，然后通过测验进一步检验这个假设。在这方面最为突出的是关于智力结构和人格结构的理论研究。不论是斯皮尔曼的智力二因素理论，还是瑟斯顿的智力群因素理论，还是吉尔福特的智力三维结构理论都是建立在对智力测验结果的因素分析基础上的。这些智力理论来源于智力测验，反过来又成为进一步编制智力测验的理论基础。在人格结构的研究中，如卡特尔的16种人格因素结构理论、艾森克的人格维度理论也都是在对人格测验结果作反复的因素分析的基础上提出来的。在教育研究中，例如要比较各种教育措施的实际效果，就需要运用教育测验获得测量分数，并对分数进行统计比较。从80年代开始，在理论研究中，有的学者特别强调非智力因素在学生学习活动中的重要作用。但在未得到实证研究结果的证明之前，这种观点只能是一种理论假设。近几年来，一些测量学工作者对这个假设进行了多方面的测验研究，发现有些非智力因

素对学生成就具有明显影响，而另一些因素的影响则不明显。这些研究为理论上的进一步探讨提供了重要的资料。

3. 实验分组

心理与教育测验还可以和实验方法结合起来运用于研究工作中。在一些实验心理学的研究课题中，为了考察不同自变量对被试因变量的不同影响，通常选择两组被试进行比较研究，这时需要控制与实验变量无关的被试的其他心理变量（例如智力水平），使两组被式实现等组化（如使他们的智力水平相当），心理与教育测验（如智力测验）可以满足实验设计中的上述要求。有时，我们需要研究具有不同心理特征的被试在完成心理实验任务过程中的差异。在这种情况下，我们首先需要通过心理测验识别不同心理特征的被试，然后分成两个极端组进行比较实验。例如，要研究内向的人和外向的人在场独立性实验中的差异，就可以先运用内外向测验选择出典型外向的被试和典型内向的被试，然后让他们完成场独立性的实验任务，从而比较他们是否在场独立性方面存在差异。

（二）实际应用功能

1. 选拔人才

在教育、企业、军事、艺术、体育、人事等部门，人们经常面临着选拔人才的问题，也就是需要识别那些最有可能获得成功的人。在传统社会里，选拔人才主要依靠少数人的经验，这显然是一种非常原始的选才方式。现代社会各行各业需要大量不同类型、不同层次的人才，那种伯乐识马式的选才方式显然不能适应现代社会对人才的需求。心理与教育测验的发展为大规模地选拔人才提供了可能。心理测量学家根据对各种工作的性质和特点的分析，寻找出适应特定工作要求的心理模式，然后根据这种模式编制测验，借此识别适合从事这种工作的

人。这就不仅大大提高了选才的效率，而且可以避免选才过程中的各种人为因素的影响，从而提高选才的科学性和客观性。美国在 1942 年第二次世界大战期间将心理测验应用于飞行员的选拔，结果淘汰率由原来的 65% 下降到 36%。心理测验在人才选拔中的价值由此可见一斑。

2. 人员安置

随着社会化大生产的发展，人事分工越来越细，不同的工种需要不同的人来做，不同的人适合做各种不同的工种，借助于心理与教育测验可以使人与事做到最佳分配，做到人尽其才，提高劳动生产率。在教育领域，可以借助于心理与教育测验的资料，作为按能力和成绩分班的依据，为分类教育、因材施教提供条件。

3. 心理诊断

对于智力缺陷者和心理障碍者的识别是推动心理测验发展的重要动力。直到现在，对各种智力落后、精神疾病和脑功能障碍应用心理测验来诊断仍然是一种重要的途径。

心理与教育测验的诊断功能不只限于临床，在教育工作中同样可以发挥作用。例如，可以应用测验发现学生学业成绩不良或社会适应不良的原因，查明学习困难或造成困难的症结所在，从而采取适当的帮助和补救措施。

4. 描述评价

应用心理与教育测验可以对人们在智力水平、学业成就、人格特点等心理特质上的优势和劣势做出描述和评价，使一个人知道自己的长处和短处，以便扬长避短，更好地学习、工作和生活。这种评价既可由他人做出，也可由自己做出；既可用于评价学生，也可用于评价教师；既可评价个人，也可评价团体。

5. 心理咨询

应用心理与教育测验获得的资料，可以作为从事心理咨询工作的依据。例如，综合成就测验、智力测验、能力倾向测验、职业兴趣测验和性格测验的资料，可以就一个人的未来职业方向提供咨询意见，以便帮助来访者做出正确的职业选择。利用人格测验和临床精神障碍测验的资料，可以帮助来访者改善心理环境，提高心理适应的能力。

第四节　心理与教育测量工作者的素质要求及道德准则

一、心理与教育测量工作者的素质要求

从前面各节的讨论中，我们可以看到，心理与教育测量工作是一项高度复杂和高度专业的工作，也是一项从理论到技术尚不很完善的工作。因此，只有不断提高心理与教育测量工作者的专业素质，才能促进心理与教育测量工作沿着科学、健康的轨道发展。改革开放以来，中国心理教育测量工作在从恢复到发展的同时，也出现了误用、滥用心理测验的现象。一些地方、一些个人随便使用心理测验，对测量结果乱加解释，引起一些不良影响。之所以出现这种情况，从根本上说，是测验的使用者缺乏应有的基本素质。因此，我们认为有必要对心理与教育测量工作者提出一定的素质要求，以便规范专业训练和有关的培训工作，培养合格的测量学工作者。

(一) 心理与教育测量工作的知识结构

概括地说，心理与教育测量工作者可分为两个不同的层次：第一个层次是专业研究工作者，主要从事心理与教育测量学的理论研究工作和各种测验的编制工作；第二个层次是实际应用工作者，主要从事运用心理与教育测验解决各行各业的实际问题的工作。不论哪个层次的人员都应具备从事测量工作基本知识结构，只是对高层次的研究人才的要求更高，应当成为该领域的专家。总结心理与教育测量发展历史的经验教训，结合中国测量学界的现状和未来发展的趋势，我们认为心理与教育测量工作者应当具备相应的基础知识和专业知识。基础知识包括：①普通心理学、发展心理学、教育心理学等广泛的心理学基本知识。②扎实的心理与教育统计学的基本知识。③教育学的基本知识。在专业知识方面，除了精通人格心理学、智力心理学、变态心理学、心理与教育测量的原理与技术等具有核心地位的专业知识外，还应根据自己的工作领域具备相应的其他专业知识。例如，在教育领域从事教育测量工作的人员应精通各个学科的专业知识；在临床领域从事心理测量工作的人员除了具备基本的医学知识外，尤其应精通精神、神经医学的专门知识；在工业企业、人事部门从事心理测量工作的人员应懂得组织人事管理知识和有关的技术知识；在司法部门从事心理测量工作的人员应懂得犯罪学、罪犯改造学、犯罪心理学的专门知识等等。总之，合理的知识结构是保证心理与教育测量工作科学化和专业化的基本条件。



(二) 对心理与教育测验的科学态度

人们对心理与教育测验的争论自测验问世以来就从未间断过。其极端的看法是要么高估测验的作用，把它奉为神明；要

么贬低测验的作用，把它视为江湖骗术。这两种态度都是极其错误的和不科学的。对此，心理与教育测量工作者应有清醒的认识。

从心理学的发展历史来看，心理测验是在心理学思辨科学转向实验科学后出现的。心理测验方法既受到心理实验方法的影响，又是对实验方法的有益补充。尤其是在研究人的较为复杂和高级的心理现象（如智力和人格）中，测验方法起到了实验方法所无法替代的作用。测验方法在客观上为心理学的发展和进步作出了重要贡献，并在众多的应用领域发挥了它的实际作用。

但是在另一方面，我们也应当看到，心理与教育测验无论在理论上还是在技术上都存在不少问题。例如，在智力测验和人格测验的编制工作中，人们首先碰到的麻烦是对什么是智力、什么是人格的问题还没有一个统一的认识。在这种情况下，测验所测量的结果究竟代表的是什么？这是一个伤人脑筋的问题。当然，这种情况在科学发展史上并不鲜见。例如，物体重量测量技术在万有引力定律被发现很早以前就被人们广泛地应用了；物体温度的测量技术在人们认识到物体分子热运动加速的原理之前也被广泛地应用了。正是由于杆秤和温度计的发明和广泛应用，才推进了对物理现象的研究，发展了物理学理论。这说明，一方面，测量技术的发展受理论研究水平的制约；另一方面，测量技术的应用反过来促进着理论研究的扩展和深化。在心理学领域，智力测验的发展深化了对智力本质及其结构的认识也是人们公认的客观事实。因此，心理与教育测量工作者一方面要认识到心理与教育测验是从事心理学与教育学研究的一种重要方法，也是解决实际应用问题的一种重要的辅助工具；另一方面也要充分考虑到目前的心理与教育测验的科学性还不够高，有待于在使用过程中进一步改进和完善。

测验起源于对个别差异的测量，但测验方法不是鉴别个别差异的唯一方法。如同心理学的任何其他研究方法一样，测验方法既有它有长处，也有它的不足。只有根据研究工作的需要将各种研究方法结合起来才能对人的心理现象获得相对全面的认识。在运用测验解决实际问题时，使用者应当记住测量结果（分数）只是对人的智力人格的相对估计，而不是一个十分精确的数值。在解释个人测量分数并以此分数为依据对个人的未来作预测时应当特别小心谨慎。

二、心理与教育测量工作者的道德准则

从事每一种职业都应遵守其特定的职业规范和道德准则。心理与教育测量工作者应自觉遵守中国心理学会于 1993 年在《心理学报》第 2 期颁布的《心理测验管理条例（试行）》和《心理测验工作者的道德准则》（见本书附录一和附录二）的规定。这里结合中国心理与教育测量的实际作些阐述。

（一）测验的保密和控制使用

心理与教育测量工具需要保密，对测验的占有范围需要控制。这本是测量学上的常识。不过，对于初学者，我们仍有必要说明保密和控制使用测验的理由。

对测验保密是为了保证测验的价值，防止测验失效。在实施测验时，人们经常碰到类似这样的情况：一所小学尝试用智力测验对新入学儿童的智力水平进行识别，以此作为分班和因材施教的参考依据。有的家长为了使自己的孩子能够进入理想的班级，找到心理测验工作者，说："让我的孩子先做做这个

测验行吗？我希望他能够测到一个好分数，能进人一个好的班级。”假如测验工作者满足了家长的要求，那么，这个测验对于鉴别这个儿童的智力水平就毫无价值了。

当然，对测验内容的保密，并不意味着不需要对受测者和一般的公众介绍关于测验的知识。但这种介绍的目的应限于：①破除对测验的神秘感；②了解测验的一些技术和方法；③熟悉测验的程序和手续，消除受测者的紧张和焦虑。

为了保证测验的保密性，《心理测验管理条例（试行）》规定：修订与出售他人所编制的心理测验时，必须首先征得该测验的主管单位或作者的同意；印制、发行与出售心理测验器材的机构应到中国心理学会心理测量专业委员会登记，并只能将测验器材售于具有测验使用资格者；为了保证测验的科学性和实用价值，标准化测验的内容与器材不得在非专业刊物上发表。《心理测验工作者的道德准则》中也规定：为维护心理测验的有效性，凡规定不宜公开的心理测验内容、器材、评分标准以及常模等，均应保密。

所谓对测验的控制使用，是指并非所有的人都可以接触和使用测验，测验的使用者必须是经过专业训练和具有一定资格的人员。对测验之所以要控制使用，是为了保证测验的实施和对测验分数的解释既做到合乎科学，又对受测者未来的成长有益。在测验工作中，人们也常碰到类似这样的情况：一个女青年愁眉苦脸地来找心理咨询工作者，诉说：“有人给我做了一个人格测验，说我的神经质分数高。此后，我就经常心神不宁，寝食不安。”显然，这是由于测验人员缺乏专业知识，对测验结果的解释不慎，给受测者造成的心理负担。这样的解释不仅无助于克服受测者的神经质倾向，而且会加重这种倾向。

为了保证对测验的控制使用，《心理测验管理条例（试行）》对测验使用人员的资格作了规定：①心理专业本科以上

学历者。②在心理测量专家的指导下，具有两年以上测验使用经验者。③经过心理测量培训班的专门训练并获得资格认定书者。其中资格认定书被分为两种：单项资格认定书和多项资格认定书。同时，《心理测验工作者的道德准则》对测验分数的解释原则也作了规定：心理测验工作者在介绍测验的效能时，必须提供真实和准确的信息，避免感情用事、虚假断言和曲解；应以正确的方式将所测结果告知被测者或有关人员，并提供有益的帮助与建议。

（二）测验中个人隐私的保护

在测验工作中，尤其是人格测验工作中经常遇到的一个不可忽视的问题是侵犯受测者的个人隐私问题。例如，在编制关于情绪、动机或态度等测验时，其中有的内容都会涉及到人们的家庭关系、内心冲突、私人生活等问题。在日常生活中，人们一般不愿意向别人透露这些事情，而在测验条件下，为了寻求帮助或配合测验，很可能表露出来。这种情况在能力测验中同样存在。因为任何智力的、能力的或成就的测验都会显示出一个人的某种缺陷， 而在一般情况下，人们是不愿意透露这些缺陷的。即使在测验条件下，受测者也会产生顾虑。在这种情况下，保护受测者的个人隐私就成为测验工作者的一项重要的责任。为此，测验工作者应当采取适当的保护措施：一是只有在必要的情况下，测验工作者才能询问个人的隐私，凡是与测验目的无关的方面就不应涉及；二是对受测者保证为其保密，并在实际上为受测者严守秘密；三是凡测验中必需涉及的个人隐私应事先征得受测者本人或其他有关人员的同意。

为了保证测验中个人隐私不受侵犯，《心理测验工作者的道德准则》规定：心理测验工作者应尊重被测者的人格，对测量中获得的个人信息要加以保密，除非对个人或社会可能造成

危害的情况，才能告知有关方面。

练习与思考

1. 与物理测量相比较，阐明心理测量的特点。

2. 为什么人们不像对物理测量那样容易接受心理测量？如何才能证明心理测量的必要性和可能性？

3*. 试结合中国心理测量界的现状讨论中国心理学会颁布《心理测验管理条例（试行）》和《心理测验工作者的道德准则》两个文件的意义。

第二章　心理与教育测量的产生与发展

本章提要：

- 中国古代学者对心理与教育测量的贡献及其特点
- 西方心理与教育测量起源的社会根源与学术背景
- 西方心理与教育测量的早期探索
- 西方心理与教育测量运动的发展
- 现代心理与教育测量在中国的发展

每一门学科都有其产生和发展的历史。了解心理与教育测量学产生和发展的历史，有助于进一步认识该学科的性质及其现状。本章将对心理与教育测量学的发展历程作一简要回顾，使读者能够对这一学科的历史有一个概要性的了解。

第一节　中国古代的心理与教育测量

和许多科学理论与科学发明均起源于古代中国一样，中国也是心理与教育测量技术的最早故乡。在漫长的中国古代社会里，现代心理与教育测量的主要领域（如能力测量、人格测量、学绩测量）均有所涉及，并取得举世公认的成就。

一、能力测量

中国古代的能力测量可上溯到2500年前的思想家和教育家孔子（公元前551～前479年），他在教育实践中凭借自己的经验观察首先评定学生能力的个别差异，并将人的智力分成三个等级，即中上之人、中人和中下之人。他说："中上之人可以语上也，中下之人不可以语上也。"[①] 用现代人的说法，就是智力水平较普通人（中人）高的人可以给以高等教育，智力水平较普通人（中人）低的人不可以给以高等教育。

① 《论语·雍也》。

汉代学者董仲舒（约公元前179～前104年）已论及到注意测验，他说："一手画方，一手画圆，莫能成。"[1] 这无疑是世界上最早的心理（注意）测验。此后，东汉王充的《论衡·书解篇》、北齐刘昼的《新论》、明代王守仁的《传习录》等都照录了这个测验。

三国时期刘邵的《人物志》可以说是一部研究能力（也包括性格，见稍后）的专门著作。在该书中，刘邵把人的才能划分为12种类型，即：清节、法家、术家、国体、器能、臧否、伎俩、智意、文章、儒学、口辩和雄杰。1937年，美国学者施瑞奥克（J. K. Shryock）将该书以《人类能力的研究》为书名译成英文在美国发表。

6世纪中叶，中国江南就有了类似于现在的婴儿发展测验的"周岁试儿"习俗。对此，颜之推在《颜氏家训》中作了详细记载："江南风俗，儿生一期，为制新衣，盥浴装饰。男则用弓矢纸笔，女则用刀尺针缕，并加饮食之物及珍宝服玩，置之儿前，观其发意所取以验贪廉、智愚，名之为试儿。"[2]

出现于清代的益智图（俗称七巧板）、九连环可以认为是最早的创造力测验。益智图用形状大小不同的七块小木块能够组合成上百种动植物和其他实物图案。九连环的设计之巧妙，足可以与现代的魔方相媲美。后来，刘湛恩用英文撰写了《中国人用的非文字智力测验》一文，将七巧板、九连环介绍到国外。美国心理学家武德沃斯（Woodworth）对九连环极为赞赏，把它视为"中国式的迷津"。至于七巧板的操作，则与现在的发散思维测验完全一致。

---

① 《春秋繁露·天无二道》。
② 《颜氏家训·风操篇》。

二、人格测量

孔子不仅论及到学生智力水平的评定，同时也提出了性格类型的观点。他说："不得中行而与之，必也狂狷乎？狂者进取，狷者有所不为也。"① 这里，孔子显然把人分成3种：狂者（以积极进取、敢作敢为为特征）、狷者（以拘谨胆怯、唯唯诺诺为特征）和中行（介于狂者与狷者之间，不偏不倚）。很明显，孔子的"狂者"相当于外倾型，"狷者"相当于内倾型，而"中行"相当于中间型。

表 2.1　刘邵编制的性格类型表②

性格类型	性格总的特征	性格的优缺点
强毅之人	狠刚不和	厉直刚毅，材在矫正，失在激许。
柔顺之人	缓心宽断	柔顺安恕，每在宽容，失在少决。
雄悍之人	气奋勇决	雄悍杰健，任在胆烈，失在多忌。
俱慎之人	畏患多忌	精良畏俱，善在恭谨，失在多疑。
凌楷之人	秉意劲特	强楷坚劲，用在桢干，失在专固。
辩博之人	论理赡给	论辩理绎，能在释结，失在流宕。
弘普之人	意爱周洽	普博周给，弘在覆裕，失在溷浊。
狷介之人	砭清激浊	清介廉洁，节在俭固，失在拘扃。
休动之人	志慕超越	休动磊落，业在攀跻，失在疏远。
沉静之人	道思回复	沉静机密，精在玄微，失在迟缓。
朴露之人	申疑实	朴露劲尽，质在中诚，失在不微。
韬谲之人	原度取容	多智韬情，权在谲略，失在依违。

刘邵根据阴阳、五行（木、金、火、土、水）和形体

① 《论语·子路》。
② 高觉敷主编：《中国心理学史》，人民教育出版社1985年出版，第177页。

（骨、筋、气、肌、血）的关系及其人的行为表现，把人的性格划分成12种类型（详见表2.1）。

三、教育测量

据迄今可考的史料分析，世界上最早的教育测量出现于中国西周奴隶制时期（公元前1100～前771年）。《礼记·学记》记载，在西周的“国学”中已经建立具有相当系统性的教育测量制度：“比年入学，中年考校。一年视离经辨志，三年视敬业乐群，五年视博习亲师，七年视论学取友，谓之小成。九年知类通达，强立而不反，谓之大成。”这一制度，不仅规定了学业考试的时间和步骤——每隔一年进行一次。而且规定了考试的内容和标准——第一年考查分析经文、章句的能力及学习志向；第三年考查学习态度及与学友的互助；第五年考查学业的广博程度及尊师情况；第七年考查分析、评价学业问题的能力，以及择善而交的能力，如果达到标准，称之为“小成”；第九年则要考查推理论事，触类旁通的能力和是否具有坚定不移的意志，是否不再有违反师长教诲的地方，如果能够达到标准，就称之为“大成”。

汉代在考试制度、考试类型和考试功能方面都作了重要的发展。在考试制度方面，调整太学考试时间，汉武帝初年曾制定了岁考制：“一岁皆辄课”，把太学的考试时间一度缩小到一年一试。在考试类型方面，开始使用3种形式考试学生：“口试”、“策试”及“射策”，首开了笔试的先河，比欧美国家早1800多年。在考试功能方面，汉代十分重视教育测量功能的发挥，已经把考试运用于督促和检查学生的学习，使考试成为

了太学学校管理的手段之一。

除了学校内部测量的自身发展以外，取士制度自汉以来的发展，客观上也对我国的古代乃至于西方的教育测量发展起了促进作用。东汉时，选拔官吏主要通过征辟和察举，乡选里举进行，至西汉开始渗入考试因素，经过魏晋南北朝时期的“九品中正”制，逐步实现了制度化，最后至隋炀帝大业二年（公元606年）发展成为科举制，在中国延续了1300年。其间，不仅创造了分科考试、“弥封”、复评等方法，而且在命题、考试组织、反舞弊等方面形成了一整套制度，不仅对欧、美的公务员制度的建立方面，而且在教育测量方面都产生了较大的影响。

四、对测验理论的最初探索

在古代中国，不仅在测验实践方面作出了杰出贡献，而且在测验理论上也有着惊人的突破。孔子之后约150年，大思想家孟子（公元前327～前289）就指出了测量人类心理的必要性和可能性。他说：“权，然后知轻重；度，然后知长短。物皆然，心为甚。”[①] 西方学者直到20世纪20年代才解决了这一理论问题。

前面提到的刘邵曾对人才鉴定的意义、可能、困难和方法作了系统的论述。刘邵认为，人才鉴定对于知人善任、振兴国家事业具有重要意义。他说：“夫圣贤之所美，莫美乎聪明。聪明之所贵，莫贵乎知人。知人诚智，则众材得其序，而庶绩

① 《孟子·梁惠王上》。

之业兴矣。"① 这里，所谓知人，就是要对人的才能和性格做出合乎客观实际的鉴定。人才所以能够鉴定是因为人的才能和性格必然要表现在人的外部行为中。他把人的行为表现概括为9种，称为"九征"。通过观察这9种外部表现就可以知道人的才能和性格的特点。即：由"神"可知"平陂之质"；由"精"可知"明暗之实"；由"筋"可知"勇怯之势"；由"骨"可知"强弱之植"；由"气"可知"躁静之决"；由"色"可知"惨怿之情"；由"仪"可知"衰正之形"；由"容"可知"态度之动"；由"言"可知"缓急之状"②。尽管如此，刘邵仍然认为人才鉴定很不容易。这主要是因为：一方面，鉴定者"各自立度"，用各自的标准去衡量人才，就很难全面地识别一个人；另一方面，被鉴定者"表里不一"、行为"似是而非"，常常令鉴定者迷惑不解。为了克服这些困难，他提出了一套识别人才的方法，即所谓"八观"和"五视"。所谓"八观"就是："一曰观其夺救，以明间杂；二曰观其感变，以审常度；三曰观其志质，以知其名；四曰观其所由，以辨依似；五曰观其所爱，以知通塞；六曰观其情机，以辨恕惑；七曰观其所短，以知所长；八曰观其聪明，以知所达。"③ 所谓"五视"就是："居，视其所安；达，视其所举；富，视其所与；穷，视其所为；贫，视其所取。"④ 这是自孔子以来对观察法的系统总结。

简单回顾中国古代社会的心理与教育测量思想，可以总结出它的几个重要特点：第一，中国古代社会的心理与教育测量思想都是描述性的，而非定量的。这当然和当时的整个科学技术水平是相适应的。第二，中国古代社会的心理与教育测量是

---

① 《人物志·序》。
② 《高觉敷主编：《中国心理学史》，人民教育出版社1985年出版，第178页。
③ 《人物志·八观》。
④ 《人物志·效难》。

分类式的。就能力测量和成就测量（即科举考试）而言，是分成高、中、低几个层次；就性格测量而言，是分成若干种类型。第三，在中国古代社会的心理与教育测量思想中注重对人作整体的鉴定和评价，并倾向于和人的道德品质联系起来。第四，中国古代社会的心理与教育测量思想与教育中的因材施教及人才使用有着密切的联系，它一开始就具强烈的应用性质。

第二节　现代心理与教育测量在西方国家的产生和发展

由于众多因素特别是由于中国人文传统的影响，中国虽然是心理与教育测量的最早故乡，但现代心理与教育测量的理论和技术不是产生于中国，而是产生于工业革命后的一些西方国家。

一、现代心理与教育测量的起源

承认人的个别差异及其对个人行为的重要影响是开展心理与教育测量工作的基本前提。在中国，2000 多年前就有了这方面的自觉探索和思考。但在西方，科学家最初发现人的心理的个别差异的重要性是起因于 18 世纪天文学上的一个偶然事件。1796 年，英国格林威治天文台的皇家天文学家 N·马斯林基因为其助手金内布鲁克观察星体通过的时间比自己晚了 0.8

秒钟，就断定他“师心自用，不依法行事”而将他辞退。20年后，另一天文学家贝塞尔对这一事件作了研究，认为这不是金内布鲁克的过错，而是一种不可避免的个人观察的误差。贝塞尔的这一发现引起了学者们对个别差异的重视和研究。但在当时并未引起心理学家的注意。

1879年，德国心理学家冯特（W. Wundt）在莱比锡大学建立了世界上第一个心理实验室。它的主要目标是要寻求人类行为的共同规律。在研究中他发现，不同被试对同一刺激的反应常常不同。研究者最初以为这是实验设计程序上的问题。经过长时间的实验才认识到，这种差异并不是偶然的错误，而是由于个人能力上的真正差异。当时的实验心理学所研究的内容主要集中在感知觉等低级心理现象上面，而对诸如能力、人格等高级心理特征还无能为力。这就为日后开展对个别差异的测量学研究提出了课题。

同时，实验心理学从一开始形成了强调严格控制实验条件的传统。这种使所有被试在尽可能标准化的条件下完成实验的传统被测验学家所继承。心理与教育测量发展到今天，测验的标准化程度已成为鉴定测验科学性的重要的指标。

促使产生心理与教育测量技术的最重要的因素是社会发展的需要。

工业革命成功后，西方国家对劳动力的需求急剧增加，工厂大量雇佣童工。为了使低能者能寻找到维生的职业，一些地方官员与工厂主订约，每雇佣20名童工，必须同时带雇1名低能者。为了设法使低能者尽可能适应工厂技术的要求，法国医生沈干（E. Seguin）开始训智力落后儿童，并于1837年创办了第一所专门教育智力落后儿童的学校。1846年，沈干出版了《白痴：用生理学方法诊断与治疗》一书，介绍了在感觉和肌肉运动方面训练智力落后儿童的方法。1848年，沈干移

居美国，将他的方法予以宣传，并得到广泛的接受。他的著作中的一些内容现已转化为能力操作测验的组成部分。

19世纪，由于科学的发展和欧洲人道主义思想的广泛传播，人们对智力缺陷者和精神病人的态度发生了重要变化，开设了一些医院专门护理和医治精神病人。这就在客观上要求确定鉴别各种心理疾病的统一标准。法国医生艾斯克罗尔（E.Esqurol）首次对智力落后与精神病作了区分，认为精神病的显著标志是情绪障碍，而智力落后的主要特征是从婴儿期就表现出来的智力缺陷。他还认为，智力落后从接近正常到最严重的白痴之间有一系列等级，而诊断智力落后程度的最可靠的方法是观察儿童运用语言的能力。他的这一思想至今还体现在智力测验之中。

随着工业技术的深刻变革，社会分工的日益精细，对劳动力能力的要求越益严格，社会上产生了对职业选拔和训练的需要，这也是促成心理与教育测验出现的因素。

二、心理与教育测量的早期探索者

（一）高尔顿

对现代心理与教育测量的产生起过直接推动作用的是英国优生学的创始人弗兰西斯·高尔顿（Francis Galton）。他是达尔文的表兄弟，深受进化论的影响。1869年，高尔顿出版了《遗传的天才》一书，提出人的能力是由遗传而来的，并设想不同人的能力水平的分布是正态的，其差异是可以测量的。1884年，高尔顿在伦敦国际博览会上成立了一个“人类测量实验室”，参观者可得到自己身高、体重、阔度等身体素质和

视听敏锐度、肌肉力量、反应时以及其他的感觉——运动机能的量化信息。博览会闭幕后，高尔顿把实验室迁到伦敦南克圣顿博物院，继续工作了长达6年之久。通过这种方法，高尔顿积累了有关简单心理现象的个别差异的大量系统的资料，这可以视为第一个大规模系统测量人的个别差异的尝试。

高尔顿在他的实验室里发明了许多测量仪器，如用于测量长度视觉辨别的高尔顿棒、用于测量听力的高尔顿笛，其中的有些仪器到现在仍然有效。他还是应用评定量表、问卷法及自由联想法的先驱。

高尔顿在心理与教育测量史上最重要的贡献之一，是把统计方法应用到对个别差异资料的分析之中。他不但扩充了百分位法，而且创造了一种简单的计算相关系数的方法。其中后者被他的学生皮尔逊（Karl Pearson）所继承和发展，创立了积差相关公式，成为当今测量学上应用最为广泛的统计工具之一。

（二）卡特尔

卡特尔（James M. Cattell）是美国心理学家，早年师从冯特。后与高尔顿有过密切交往，并受到后者的影响。回到美国后，卡特尔致力于心理实验室的建立与测量思想的传播。1890年，卡特尔在《心理》杂志发表《心理测验与测量》一文。在这篇论文中，首次提出了“心理测验”（mental test）这个术语，并报导了他所编制的一套能力测验在大学生身上的应用结果。测验内容包括肌肉力量、视听敏感度、运动速度、重量辨别、反应力、记忆力，以及类似的一些项目。在该文中，卡特尔还论述了测验理论上的一些问题。他认为，心理学只有立足于实验与测量，才能达到如同自然科学的准确性；心理测验只有建立普遍的统一标准，并要与常模相比较，才能充分地

实现其科学价值和实用价值。这些观点都已成为测量学上的重要观念。

（三）比内

比内（A. Binet）青年时代学医，但对心理学产生兴趣。1886年出版他的第一部著作《推理心理学》，1889年与亨利·博尼（Henri Beaunis）在索那建立法国第一个心理实验室，1891年出版《人格心理学》一书，1895年创办法国心理学杂志《心理学年报》，同年与亨利（V. Henri）联名发表文章，批评当时流行的测验偏重于简单感觉，不能测出真正的智力。这种批评是正确的。因为卡特尔将他编制的测验用于施测哥伦比亚大学学生，然后计算测验分数与其考试成绩的相关，结果相关值很低。1893年，贾斯特罗（J. Jastrow）编制出一套由15个分测验组成的测验，但使用的结果却不能使人满意。这些测验的结果不仅彼此相关不高，而且与教师对学生智力水平的评价结果也没有什么相关，与学生的学业成绩的相关也不高。

比内认为，测量比较复杂的心理功能，不必苛求精确度，因为这些功能的个别差异较大。1898年，比内在哲学杂志上发表《人格心理学中的测量》一文，提到许多测验，如画方形，比较线的长短，记忆数目，词句重组，折纸，理解文章意义等。其中的许多被后来的量表所采用。在该文里，比内还提出心理测量的根本原理在于将个人的行为与他人作比较。这个观点已成为现代心理与教育测量的一个普遍原理。1903年，比内的另一部著作《智力的实验研究》问世，提出了智力的定义，认为智力是高级心理过程，包括推理、判断以及动用已知知识解决新问题的能力。他以自己的两个女儿为被试，进行词语填充、图片解释等项目的测量。这些项目也被吸收到他后来

的量表中。

1904年，一个偶然的机会使比内的思想得以实践，并由此推动了心理与教育测量的迅速发展。这一年，法国公共教育部决定成立一个有医学家、科学家和教育家组成的委员会，专门研究公立学校中落后儿童的教育方法。比内作为该委员会的成员，主张用测验方法来识别智力落后儿童，但遭到许多其他委员的反对。比内不顾众人反对，与其助手西蒙（T.Simon）合作完成了世界上第一个智力测验量表——比内－西蒙量表(Binet－Simon Scale)。1905年，他们在《心理学年报》上发表的《诊断异常儿童的新方法》一文介绍了该量表，史称1905年量表。

1905年量表由30个由易到难排列的项目组成，可用来测量各种能力，特别是判断、理解和推理能力，亦即他所谓的智力的基本组成部分。虽然其中也包含了部分感知觉的测验，但主要是语言理解测验。

1908年，比内发表了修订后的比内－西蒙智力量表，删掉了1905年量表中不合适的项目，增加了一些新的项目，使总题数增加到59个。所有项目都按年龄分组，组别从3岁～13岁。年龄水平根据300名正常儿童的测验结果确定。测验成绩用“智力水平”表示，目的在于确定儿童能够完成何种年龄水平的儿童能完成的测验，并建立了常模。

1911年发表了第二次修订本。这次修订没有重大的变化，只是改变了几种年龄水平分组，并将测验扩展到成人。就在这一年，比内逝世，终年54岁。

回顾西方心理与教育测量早期探索的历史可以看出，心理与教育测量的产生既有着深刻的社会时代背景，又与科学技术的发展水平紧密联系，同时也与科学家个人的学术贡献有关。正如美国著名心理史学家波林（E.G.Boring）所指出的，在

测验领域中，“19世纪80年代是高尔顿的10年，90年代是卡特尔10年，20世纪头10年则是比内的10年。”[①]

三、心理与教育测量运动的发展

从20世纪初叶开始，西方心理与教育测量获得迅速发展。其发展的基本轨迹是：20年代进入狂热期，40年代达到顶峰，50年代以后经典测量理论趋于成熟并稳步发展，60年代以后测量理论出现新的动向，尤其是项目反应理论和概化理论的出现引起了心理与教育测量领域的深刻变革。下面分4个方面作些简要回顾。

（一）智力测验的发展

比内－西蒙智力量表发表后，引起世界各地的广泛关注。各种语言的版本纷纷出现，其中最为著名的是美国斯坦福大学推孟（L. M. Terman）于1916年修订的斯坦福－比内量表。其中影响最为深远的变动是推孟采用了比率智商的概念来表示智力水平的高低。早在1911年，德国汉堡大学心理学家斯腾就曾提出用儿童的心理年龄与实足年龄的比值（心理商数）来表示儿童的聪明程度，推孟在修订比内－西蒙量表时将其改为“智商”，从此智商一词风靡全世界。

比内－西蒙量表及其修订形式都是个别测验，一次只能测量一个被试。这种测验在临床诊断和个案资料的收集中是有价值的，但若是测量对象太多，就非常费时间。针对这种情况，

① 转引自郑日昌：《心理测量》，湖南教育出版社1987年出版，第10页

适合于大规模测量的团体测验被发展起来。1917年，美国政府决定参加第一次世界大战。美国心理学会组成以叶克斯(R.M.Yerkes)为首的委员会讨论心理学如何为战争服务的问题。他们认为，军队在选拔和分派官兵时，应当考虑他们的智力水平。但军队有100多万人，若要实施智力测验，就只能采用团体施测方法。于是出现了“陆军甲种测验”和“陆军乙种测验”，前者为文字测验，后者为非文字测验。两种测验均可用于大规模的团体施测。在1917～1919年间，运用这两种测验共测量了200多万名官兵，积累了大量的资料。

战后，这两种测验在修订后被广泛运用到整个社会，为教育和工商人事服务。在20年代，智力测验运动出现了狂热的势头，大量的团体智力测验不断涌现，以至出现了粗制滥造的情况。

随着智力测验的发展和统计学的进步，对智力本质及其结构的统计学研究应运而生。英国心理学家斯皮尔曼(C.Spearman)首先运用因素分析方法研究智力结构，提出智力结构的“二因素理论”，推动了30年代～50年代的智力结构研究，并为编制新的智力测验奠定了理论基础。

为了满足社会对测验的需要，新的智力测验不断编制出来。30年代以后，英国心理学家瑞文(J. C. Raven)针对斯皮尔曼的“G”因素相继编制了“瑞文标准推理测验”、“瑞文彩色推理测验”、“瑞文高级推理测验”。从40年代末开始，美国心理学家韦克斯勒(D.Wechsler)也相继编制了“韦氏儿童智力量表”(1949)、“韦氏成人智力量表”(1955)和“韦氏幼儿智力量表”(1967)。韦克斯勒在智力测验方面的最重要的贡献是：①他舍弃了比率智商，而用离差智商代之，从而克服了比率智商的局限。②他编制的智力量表分为言语量表和操作量表两部分，不仅能够获得总体智力水平的信息，而且可以获得

受测者智力优势的信息。

（二）能力倾向测验的发展

智力测验所测量的只是人的一般能力水平，只是人的能力结构中的一个方面。从 20 年代开始，人们在开发智力测验的同时，着手编制特殊能力测验。最初被称为“学业能力倾向测验”，后来进一步扩展到职业咨询、工业部门及军事领域的人才选拔和安置工作。这些测验包括音乐、文书、机械和艺术等强调特殊能力的领域。在编制成套能力倾向测验的过程中，因素分析方法起了重要的作用。因为这种方法能够通过对测验的分析获得相对独立的能力因素，如言语理解、数学推理、空间定向、知觉速度、机械操作等。因此，根据因素分析法编制的测验通常提供的是被试在各个能力因素上的分数，这就有助于进行个体内部心理结构的分析。

（三）成就测验的发展

心理测量原理和技术的发展，为学校考试制度的改革提供了理论依据和技术手段。早在 1897 年，赖斯（G.Rice）就曾编制出美国学校儿童拼读能力测验。20 世纪初，桑代克（E.L.Thorndike）编制了第一个标准化的教育成就测验，该测验运用心理测量原理，编制出评定学生书写、作文、拼读、算术、计算和推理的量表。正因为如此，桑代克被公推为教育测量的鼻祖。1923 年，凯利（L.Kelley）、鲁奇（G.Ruch）和推孟合作编制了第一个成套成就测验——“斯坦福成就测验”。该测验的一个显著特点是能够对不同学科的测验成绩进行比较。由于成就测验属于客观测验，传统的论文式考试开始引起争议，认为论文式考试费时多，评分结果不可靠。

30 年代后期，在美国出现了跨州、跨区域乃至全国的测

验机构。其中最为著名的要数“大学入学考试委员会”（College Entrance Examination Board，简称 CEEB)。1947 年，美国成立“教育测验服务中心”（Educational Testing Service，简称 ETS)，它的任务是编制各种测验程序，供各大学、学校和政府机构选用。1959 年，美国又建立了“美国大学测验系统”(American College Testing Program)，该机构提供选拔获取奖学金的高材生的测量方法。

现在，成就测验不仅用于教育领域，而且被广泛地应用于工业企业的人事任用和政府公务员的选拔。

(四) 人格测验的发展

心理与教育测量的另一重要领域，是对人的人格特质的测量。这一领域涉及广泛的方面，如情绪、动机、兴趣、态度、气质、性格等等。

最早进行人格测量的是克雷培林（E. Kraepelin)，他最早用自由联想法诊断精神病人。在这样的测验中，主试给被试提供若干经过选择的刺激词，要求被试用最快的速度报告他想到的第一个词。克雷培林还用这种方法研究了疲劳、饥饿、药物的心理效应，发现所有这些状态都增加了病人的表层联想。此后，自由联想技术一直是用来诊断人格障碍的一种方法。

本世纪初叶，出现了自陈人格问卷。1917 年，美国心理学家武德沃斯用自陈问卷法编制了适用于诊断士兵神经症的“个人资料调查表”。后来，美国的卡特尔（R.B.Cattell）经过多年的努力，编制成“卡特尔 16 种人格问卷”。英国的艾森克(H.J.Eysenck）编制成“艾森克人格问卷”。美国明尼苏达大学的哈兹威（S.R.Hathawag）和莫肯利（J.C.Mckinley）编制成“明尼苏达多项人格调查表”。这些人格问卷后来被翻译成多种文字，流行于全世界。

人格测量的另一种重要的技术是投射测验。早在15世纪就有人注意到墨迹可以刺激人的想象。比内也曾想利用墨迹来测量儿童的智力，但没有成功。1910年，瑞士精神医学家罗夏克（H.Rorschach）为了研究精神障碍对知觉的影响，曾用一些画片来测量病人，以后改用墨迹图。在最初制作墨迹图时，先在一张纸的中央倒一堆墨汁，然后将纸对折挤压，使墨汁向四面流动，形成两边对称但形状不定的图形。罗夏克以此类图形，对各种精神病患者作了大量试验，发现不同类型的病人，对墨迹图有不同的反应。然后再和低能者、正常人、艺术家等的反应作比较，最后确定其中10张墨迹图作为测验材料，逐步确定记分方法和解释测验结果的原则，于1921年正式发表。此后，哈罗尔（Harrower）在第二次世界大战期间编制了以团体方式实施的墨迹测验；霍滋曼（Holtzman）也编制了墨迹测验，且有复本，每套由45张墨迹图组成。此外，1935年，由莫瑞（H.A.Marrau）和摩根（Morgan）编制的著名的"主题统觉测验"（Thematic Apperception Test，简称TAT）也是投射测验的一种。其他如句子完成测验、情境对话测验、画人测验等也属于投射测验。

四、心理与教育测量的当代趋势

60年代以后，心理与教育测量学界出现了一些新的方向。概括起来主要是3个方面：一是由于信息加工心理学的兴起，测量学界倾向于将实验法和测验法相结合，产生了信息加工测验。二是由于计算机技术的迅速发展，传统的纸笔测验逐渐被电脑程序测验所取代，从而大大提高了测验的效率。三是针对

经典测量理论（即真分数理论）的某些缺陷，提出了一些新的测量理论，尤其是项目反应理论和概化理论，不仅在理论上取得了巨大成就，而且在应用上也显示出强大的生命力。

第三节　现代心理与教育测量在中国的发展

一、现代心理与教育测量建国前的发展

清朝末年，西方心理学开始传入中国。1914 年，有人在广东对 500 名儿童作了记忆的比喻理解测验。1917 年，樊炳清首先向国人介绍了比内－西蒙智力量表。1918 年，俞子夷编制“小学生毛笔书法量表”可视为我国最早的新式教育测验。1920 年，廖世承和陈鹤琴在南京高等师范学校率先开设心理测验课程。1921 年，廖、陈二人出版《心理测验法》。1921 年，费培杰将比内量表译成中文，并在江苏、浙江二省的小学生中进行过测验。同年，中华教育改进社邀请美国测量学家麦柯尔（W.A.Mocall）来华讲学，并指导北京师范大学、北京大学、燕京大学、北京女子高等师范大学、东南大学的师生编制测验，各地编成测验 40 多种。麦柯尔评价当时中国心理学家所编制的测验“至少都与美国的水平相当，有许多竟比美国的为优。”1923 年，中华教育改进社对全国 22 个城市和 11 个乡填的 9.2 万名小学生进行了测验，引起教育界的关注。1931 年，在艾伟、陆志伟、陈鹤琴、肖孝荣等人的倡议下成立了中国测验学会。次年，《测验》杂志创刊。从 20 年代初至

40年代末，除抗战期间外，中国的心理与教育测量工作从未间断过，并涉及广泛的领域。

在智力测验方面，1924年，陆志伟根据中国南方的测验结果发表了《订正比内西蒙智力测验说明书》；1936年，他和吴天敏合作，将测验范围扩大到北方，作了第二次修订。他们的研究结果表明，中国儿童的智力测验成绩显著高于欧美和日本同年龄的儿童。在此期间，廖世承编制了"团体智力测验"，陈鹤琴编制了"图形智力测验"，刘湛恩编制了"非文字力智力测验"，均有一定影响。

在人格测验方面，肖孝荣曾修订了"武德沃斯个人资料记录表"，并编制有9～15岁的常模。1935年，浙江的沈有乾用"朋洛德人格问卷"测量中国学生，发现中国男生的神经症倾向明显高于美国男生。1937年，周先庚用"塞斯顿情绪稳定性测验"测量中国学生，也发现中国学生的情绪适应性较差。1943年，林传鼎试用"普莱西X－0测验"，发现中国11～18岁青少年的情绪成熟度的发育比美国同年龄青少年晚一年左右。1948年，刘范曾试用"罗夏克墨迹测验"。

在教育测验方面，艾伟曾编制小学儿童各科学绩测验10多种。特别是他对中学生阅读能力和理解能力的研究，为当时的语文教学改革提供了科学依据。

这时期在测量学学科建设方面，共出版有关心理与教育测量的著作达20余种，其中孟承先的《测验之学理的研究》、王征葵的《态度测量法》、沈有乾的《心理与测验》、王书林的《心理与教育测验》、陈选善的《教育测验》、艾伟的《小学儿童能力测验》以及孙正邦的《心理与教育测验》等具有较大的影响。

二、现代心理与教育测量建国后的发展

1949年以后的30年间，由于深受前苏联心理学的影响[①]，我国心理与教育测量一直是一个禁区，无人问津。1979年，随着心理科学在中国现代化进程中的地位得到重新肯定，心理与教育测量工作也得以恢复。10多年来，中国的心理与教育测量在各主要领域获得了迅速发展。为心理与教育测量的学科建设和改进实际工作作出了重要贡献。

（一）智力测验

1979年，中国心理学会医学心理专业委员会在天津成立心理测验协作组，决定由龚耀先主持修订“韦克斯勒成人智力量表”，于1982年完成修订工作。1980年，中国心理学会实验心理学专业委员会在武汉成立心理测验协作组，决定由林传鼎、张厚粲主持修订“韦克斯勒儿童智力量表”，于1986年完成修订工作。1982年，吴天敏修订出版了“第三次修订中国比内测验”。1986年，龚耀先又主持修订了“韦克斯勒幼儿智力量表”。1985年，张厚粲主持修订了“瑞文标准推理测验”。1989年，李丹主持修订成“瑞文测验（联合型）”。1992年，戴忠恒修订了“一般能力倾向测验”。这些量表都是国际上著名的智力测验，修订后广泛用于智力问题的研究和因材施教、人才选拔、职业咨询、临床诊断等领域。此外，中国学者也编

① 1936年，苏联开展对“儿童学”的批判运动。受扩大化的影响，心理与教育测量成为研究禁区。

制了不少的智力测验，如张厚粲、周容等编制的“中国儿童发展量表”等等。

（二）人格测验

1982年，由宋维真主持修订“明尼苏达多相人格调查表”，于1985年完成修订工作。1981年，刘绍衣等修订了“卡特尔16种人格因素问卷”，制定了辽宁省的常模，后于1988年由戴忠恒、祝蓓里主持制定出全国常模。1983年，龚耀先主持修订了“艾森克人格问卷”，同时陈仲庚也在北方作了修订。在临床心理学领域，也修订了多种涉及心理健康评估的量表，如张明园于1987年修订了“生活事件量表”。吴文源等于1990年修订了“症状自评量表（SCL－90）”等等。

80年代末90年代初，中国心理学家在继续引进修订国外人格量表的同时，开始编制中国人自己的人格量表。1988年，洪德厚等编制了“中国少年非智力个性特征问卷”。1992年，宋维真等人在借鉴“明尼苏达多相人格调查表”的基础上，编制出“心理健康测查表”，1993年，他们同香港学者合作编制出“中国人个性测量表”。1992年，沙毓英、张锋等人编制出“学生性格量表（11～18岁）”，并于1995年在张锋的主持下制定了云南省城市、农村及少数民族学生的常模。

（三）教育测验

1979年，林传鼎、张厚粲等人在参考国外资料的基础上编制了“少年儿童学习能力测验”，用于测量小学毕业生的普通能力，并估计小学毕业生是否具备学习初中课程所必需的语言能力和推理能力。

1980年开始，在张厚粲的主持下，北京师范大学高考研究组对每年的高考试卷作了系统的统计分析，获得了有关试卷

信度、效度、难度、区分度等有意义的信息。他们还对高考试卷评分的客观性、考试科目的合理设置及各科分数的合理组合作了研究。在此基础上对我国高考制度的进一步改革提出了一系列重要的意见和建议。

1984年，我国正式加入世界上最有影响力的“国际教育成就评价协会”（IEA），并与“国际教育成就评价协会”合作，在我国进行了全国规模的教育测量抽样研究。

1986年，罗黎辉、施良方等对教育目标分类理论进行了研究，并将50年代以来给国际教育测量学研究带来突破性进展的重要理论即布卢姆（B.S.Bloom）的《教育目标分类学》系统地介绍到中国，为我国教育测量的理论研究与实践提供了新的视野，注入了新的气息，产生了广泛而积极的影响。

80年代，张敏强、张厚粲对经典测量理论和项目反应理论在考试制度改革中的应用情况作了比较研究。杨志明、张厚粲运用概化理论对测量误差作了分析。张厚粲等人以项目反应理论为基础建立了“普通心理学计算机化适应性测验系统”。漆书青、戴海崎等人以项目反应理论为依据编制了“党务工作者专业知识计算机化自适应测验”，为干部考核的科学化迈出了重要的一步。胡显勇运用概化理论对作文评分误差的控制作了研究。

标准化考试理论与实践的研究、题库理论与技术的研究取得长足进展。《标准化考试简介》（国家教育委员会学生管理司，1985）、《标准化考试的理论与实践》（廖平胜等，1986）、《题库建设的理论与实践》（国家教育委员会考试中心，1991）等一大批著作相继出版。1985年国家教育委员会开始在广东省进行了高考标准化的试验。1989年，华南理工大学建设的“高等数学试题库”及“高校工科物理题库”通过国家教委鉴定，全国各地大、中、小学各学科的各种题库纷纷建成并投入

使用。

80年代以来，教育测量开始从单一的学生学绩测量，逐步向多侧面发展。学生发展测量，教师教学质量测量、课程建设质量测量、办学效益测量等各种教育测量悄然兴起。教育测量类型也从过去比较单一的终结性测量发展为诊断性测量、形成性测量等多类型的测量。

（四）组织建设和人才培养

随着心理与教育测量研究的深入和应用领域的拓展，心理与教育测量的组织建设也得到强化。1984年，中国心理学会组建心理测验工作委员会，后进一步扩建为心理测量专业委员会。该专业委员会定期组织召开全国性的学术会议。针对近年来滥用和误用测验的情况，该专业委员会及时制定了《心理测验管理条例》和《心理测验工作者的道德准则》两个文件，由《心理学报》公开颁布。此外，教育学界也成立了教育统计与测量学会，开展有关工作。

保证心理与教育测量事业健康发展的前提条件是培养合格的测量学人才。继1980年北京师范大学心理系率先开设“心理测量”课程以来，各大学有关系科都已先后开设了“心理与教育测量”课程。一些大学招收硕士和博士研究生，为该学科培养高层次学术人才。心理测量专业委员会以及一些大学还多次举办心理测量技术培训班，培养了一批应用型人才。

（五）学科建设

80年代以来，国内不仅发表了大量有关心理与教育测量的研究论文，而且出版了多种教材与专著。如宋兆鸿等的《现代教育测量》(1986)，郑日昌的《心理测量》(1987)，戴忠恒的《教育与心理测量》(1987)，余嘉元的《教育与心理测量》

(1987)，王汉澜主编的《教育测量学》(1988)，彭凯平的《心理测验——原理与实践》(1989)，邢最智、司徒伟成的《现代教育测量理论》(1989)，漆青书、戴海崎的《项目反应理论及其应用研究》(1992)，黄光扬的《心理测量的理论与应用》(1997)等等。

练习与思考

1. 简述古代中国对心理与教育测量的贡献及其特点。

2. 心理与教育测量从卡特尔到比内的发展过程中可以看出什么特点?

3. 推孟和韦克斯勒在智力测验发展过程中各有什么贡献?

4*. 比较心理与教育测量在中国解放前后的发展，会得出什么结论?

5*. 根据本章的论述，查阅有关资料，讨论中国心理与教育测量的现状与未来发展的方向。

第三章　经典测验理论的基本假设

本章提要：

●心理特质及其可测性

●心理测量的误差及其种类

●真分数的含义

●经典测验理论的基本假设

在日常生活中，人的身高、体重等特征比较容易测量。因为人的这些生理属性比较稳定而直观，其测量工具（尺子和秤等）也容易制作和使用。然而，人的内隐的心理特征是否稳定？它们能够测量吗？如果能够测量，又必须具备哪些条件？在本章里，我们先讨论经典测验理论（Classical Test Theory，CTT）的若干基本假设。

第一节　心理特质及其可测性假设

一、心理特质的含义

在日常生活中，我们发现有的人比较热情，有的人比较冷漠；有的人比较聪明，有的人比较愚笨；有的人比较急躁，有的人比较文静等等。为研究方便，我们称这种表现在一个人身上所特有的相对稳定的行为方式为人的心理特质（trait)。对这一概念，我们可以从以下几个方面来理解：

(1) 特质是一组具有内部相关的行为的概括，具有一定的抽象性。例如：某人在公共汽车上总是给老、弱 、病、残和小孩让座，在生活中总是能对他人友好相待，热情相助等等，则可以称该人具有“善良”的特质，因为在他身上总是表现出一组具有内部相关的行为（让座、友好待人、热情助人)，这种行为经概括后便具有抽象性了。如果一个人能在各种测验中获得好成绩，在工作和生活中总能想出好主意解决难题，则该人具有“聪明”的特质。

（2）特质是“一种一般的神经心理系统，……它可以综合不同的刺激，使人对这些刺激做出相同的反应”（G. Allport）。例如，某人在公共汽车上如果只给熟人和朋友让座而不理睬不认识的老、弱、病、残，则不能说他具有“善良”的特质。因为，“善良”特质要求他对各种不同的刺激（老、弱、病、残）都能做出相同的反应（让座）。

（3）特质是一个人身上比较稳定的特点。人的心理活动是十分丰富的，并不是他的每一种心理活动都会表现为一种特质，而是那些经常出现的比较稳定的心理特征才称得上特质。人们常说，“智者千虑，必有一失”，但我们在评价他时并不会因他一时之失而否认他是个智者。例如，诸葛亮尽管也吃过败仗，但我们仍然认为他是个智者。

（4）一个人的精神面貌（人格）是由多种特质分多个层次有机组合而成的。不同的人往往具有不同的特质组合，即使其特质类型相同，其特质水平往往也会有高低之分（尽管水平的高低只具有相对意义）。心理学家在研究人的人格特征时，一般是把它们分解成多个单元（特质）和层次进行分析的，并认为，人格就是多个特质多种层次的有机组合。心理测量的任务就是要区别出不同个体在能力、个性等特质上的差异。

（5）特质可以决定一个人对特定刺激的反应倾向，可以对人的行为进行某种预测。心理测量的最终目的就是要了解人的特点，并对人的行为倾向作出预测。

二、心理特质的可测性

心理特质是一种客观存在，“凡客观存在的事物都有其数

量”（E.L.Thorndike），“凡有数量的东西都可以测量”（W.A.McCall）。这就是CTT的心理特质的可测性假设。

事实上，心理特质是一种相对稳定的东西，我们可以有许多办法对它进行定义，也可以通过特殊的测量工具对它进行测量。比如，关于人的智力，目前已有了比较好的测量办法，关于人的个性，其测量方法也正逐渐成熟，并在发挥着重要作用。

当然，心理测量没有物理测量那样容易。因为人的心理特质具有比较隐蔽的特性，我们无法直接对它进行测量，只能通过被试对一些刺激（如考题）的行为反应（考试答案等）特点来推测其心理特质的特点和水平。此外，心理测量的工具也不易制作，其使用方法也比较麻烦，这都给心理测量工作增加了难度。

第二节　测量误差及其来源

一、测量误差的含义

测量误差指的是在测量过程中由那些与测量目的无关的变化因素所产生的一种不准确或不一致的测量效应。这里，我们要从两方面来进行理解：其一，测量误差是由那些与测量目的无关的变因所致；其二，测量误差表现为不准确或不一致两种方式。

例如，当我们去小摊上买水果时，若摊主偷换了秤砣，其

实测结果一定不准（误差的表现方式之一）。假若摊主的秤是合乎要求的，但他操作时故意快速地耍些手法，则其测量结果一定会与你复秤时所得结果不一致（误差的表现方式之二）。这里，误差的产生全是由那些与测量目的无关的变因（修改测量工具、不正确地使用工具）所致。

二、测量误差的种类

和物理测量一样，心理测量也有两种误差，即随机误差和系统误差。所谓随机误差即是那种由与测量目的无关的、偶然因素引起的、而又不易控制的误差。它使多次测量产生了不一致的结果，其方向和大小的变化完全是随机的，只符合某种统计规律。例如，在进行手枪射击时，新手往往很难控制手臂的轻微摆动，结果多次射击的成绩很不一致，造成误差，这种误差就是随机误差。

所谓系统误差即是那种由与测量目的无关的变因引起的一种恒定而有规律的效应。这种误差稳定地存在于每一次测量之中，此时尽管多次测量的结果非常一致，但实测结果仍与真实数值有所差异，是不正确的。例如，在射击过程中，尽管射击手非常优秀，每次结果都很一致，但若是枪的准心有点毛病，则其射击结果仍将会有稳定的偏差。又如，在进行数学测验时，若有一道10分的题的标准答案给错，则全体正确作答该题的考生的成绩将普遍下降10分，这也是系统误差。若是教师评分标准宽严不一，甚至是随心所欲，则考生的成绩还会出现较大的随机误差。

由上可知，系统误差只影响测量的准确性、不影响稳定

性。而随机误差既影响稳定性又影响准确性。

三、测量误差的来源

在物理测量中，误差来源主要有3个，即测量工具、被测对象以及施测过程。当被测对象本身不稳定或测量工具不科学，或施测时的条件、操作等不合要求时，测量便必然会出现误差。同样，心理测量的误差也来自3个方面，即测量工具、被测对象和施测过程。

在测量工具方面，心理测量与物理测量有所不同。心理测量工具通常是一套以测验（问卷）为核心的刺激反应系统（通常称作量表)。当量表在测查人的某种心理特质时，若项目所测的东西与我们欲测的目的之间出现偏差（如项目取样太少或太偏），则测量会出现误差。例如，当语文考试出现偏题时，押中题的人就会得到好成绩，没押中题的人则得不到好成绩，无法反应各人的真实水平。又如，数学测验的好坏若取决于文字理解能力的高低，则该测量也会出现误差。当一个量表对同一批人前后几次测查结果极不一致时，则认为该量表缺乏足够的稳定性。心理测量量表是否稳定、是否真正测到了我们所要测的东西是测量工具造成误差的两种主要原因。

在测量对象方面，造成测量误差的主要原因是受测者真正水平是否得到正常发挥。一般地，受测者的某种心理特质水平是相对稳定的，但是他在接受测量时的生理和心理状态会影响其水平的正常发挥。比如，当受测者过分疲劳，或突然生病、或过分焦虑、紧张时其测量成绩会低于其真实水平。如果他在进行测量的技能技巧方面经验不足，也同样会出现测量误差。

此外，受测者应试动机的强弱、受训时间的长短、受训内容的多少、答题反应的快慢等等都会产生测量误差。

在施测过程方面，产生测量误差的原因主要是一些偶然因素（恒定因素较易控制）。比如，在物理环境方面：施测现场的温度、光线、声音、桌面好坏、空间阔窄等等会造成误差。在主试者方面：主试者的年龄、性别、外表及其施测时的言谈举止、表情动作、是否按规定实施测验等等也都会造成误差。此外，评分记分环节也是容易出现差错的地方。还有，若是出现意外干扰（如：考场突然停电、有人作弊、计时表停了、试卷印刷或装订出错等），则同样会让考生分心或造成考场混乱，导致测量误差。

第三节　真分数及其有关的假设

一、真分数的含义

人的心理特质水平经测量之后应表现为一个数值。然而，由于测量误差的存在，实际测得的数值往往难以和该特质的真正水平值完全一致，它总会略高于或略低于其真实水平值，某些时候还会严重偏离其真实水平值。例如，我们平常所说的“××考生基本上考出了其应有水平”或“××被试的人格特点被基本上测出来了”或“××人这次测验超水平发挥”等等，就是对这种测量现象的一种描述。为研究方便，我们把反映被试某种心理特质真正水平的那个数值称作该特质的真分数

(True Score，简称为T分数)，把实测的分数称作该特质的观察分数(Observed Score)。当观察分数接近真分数时，就说这次测量的误差较小。

显然，真分数是一个在理论上构想出来的抽象概念，在实际测量中是很难得到的。因为任何一种测量，无论它有多么科学，总会存在误差。我们只能通过改进测量工具、完善操作方法等办法来使观察值尽量接近真分数。只要观察分数与真分数之间的误差不是太大，或者说误差被控制在可接受的范围之内，我们的测量也就可以看作是可接受的测量了。

二、数学模型及其假设

既然观察分数很难等于真分数，那二者之间是个什么关系呢？经典测验理论假定，观察分数（记为X）与真分数（T）之间是一种线性关系，并只相差一个随机误差（记之为E）。

即： $X = T + E$ (3.1)

这就是CTT的数学模型。

根据这一模型，我们可以引伸出3个相关联的假设公理(Gulliksen，1950)：

①若一个人的某种心理特质可以用平行的测验反复测量足够多次，则其观察分数的平均值会接近于真分数。

即： $E(X) = T$ 或 $E(E) = 0$

②真分数和误差分数之间的相关为零。

即：$\rho(T, E) = 0$

③各平行测验上的误差分数之间相关为零。

即：$\rho(E_1, E_2)=0$

其中，第②、第条③假设意在说明 E 是个随机误差，没有包含系统误差在内，第①条假设则在于说明 E 是个服从均值为零的正态分布的随机变量。

对 CTT 的这一数学模型及其假设公理，我们可以从以下 3 个方面来加以理解。首先，在问题的研究范围之内，反映个体某种心理特质水平的真分数是假定不会变的，测量的任务就是估计这一真分数的大小；其次，观察分数被假定等于真分数与误差分数之和。即，假定观察分数与真分数之间是线性关系，而不是其他关系；第三，测量误差是完全随机的，并服从均值为零的正态分布。它不仅独立于所测特质真分数，而且独立于所测特质以外的其他任何变量，这就保证了误差 E 中不含有系统误差成分。此外，各平行测验上误差分数间的相互独立也进一步保证了 E 的随机性，使得观察分的均值可以稳定地趋于真分数。

值得注意的是，模型假设中所提到的平行测验是个重要的概念。CTT 认为：如果两个题目不同的测验测的是同一特质，并且题目形式、数量、难度、区分度以及测查等值团体后所得分数的分布（$\bar{X}$ 和 S）都是一致的，则这两个测验被称作是彼此平行的测验。

不过，用许多个彼此平行的测验反复测量同一个人的同一种心理特质的做法往往是很难实现的，因此，CTT 的模型及假设只是一种理论上的描述。然而，有了这一模型和假设之后，却能帮助我们解决测验中的许多实际问题。

事实上，我们在实施一个标准化测验时，并不是用许多平行测验来反复测查同一批被试，而是用一个测验来同时测查许多被试。由于每个人的误差都是随机的，且服从均值为零的正态分布，所以，当被试团体足够大时，团体内的各种随机误差

会相互抵消，整个团体的观察分数的均值会趋近于该团体真分数的均值。这里，多个被试接受同一个测验相当于多个平行测验反复测查一个具有团体真分数均值水平的一个个体。因此，CTT 的理论模型和假设便派上了用场。

根据 CTT 模型和假设，我们很容易推导出如下关系：

$$S_X^2 = S_T^2 + S_E^2 \tag{3.2}$$

即：在一次测量中，被试观察分数的方差等于其真分数方差与误差分数方差之和。

注意，公式（3.2）中只涉及到了随机误差的变异，系统误差的变异包含在真分数的变异之中。即，真分数还可以分成两部分：与测量目的有关变异（S_V^2）和与测量目的无关的变异（S_I^2），即：

$$S_T^2 = S_V^2 + S_I^2 \tag{3.3}$$

于是（3.2）可改写成：

$$S_X^2 = S_V^2 S_I^2 + S_E^2 \tag{3.4}$$

这就是说，一次测验中，一个团体的实测分数之间的变异性是由与测量目的有关的变异数（S_V^2）、稳定的但出自无关来源的变异数（S_I^2）和测量误差的变异数（S_E^2）所决定的。

练习与思考

1. 简述“心理特质”的含义。
2. 心理测量的误差来源主要包括哪些？
3. CTT 模型及其假设的主要内容是什么？
4*. 根据 CTT 的数学模型及其假设，推导关系式（3.2）。

第四章　测量信度

本章提要：

- ●测量信度的概念及作用
- ●信度的估计方法
- ●影响信度的主要因素
- ●提高信度的常用方法

在各种测量活动中，常常可以看到测量者进行复测的行为。如果两次所得的测量结果比较一致，则测量者就会认定此测值；如果两次所测结果相当不一致，测量者就不敢贸然认定其中的任何一个量值。同样，在心理测量工作中，测量的结果也必须是经得起“复测”检验的。倘若不同次测量的结果有较大的差异，则这种测量的结果是难以让人信服的。本章所讨论的中心话题便是测量结果的稳定性问题，即测量的信度（reliability）问题。

第一节 信度概述

一、什么是信度

信度（reliability）指的是测量结果的稳定性程度。换句话说，若能用同一测量工具反复测量某人的同一种心理特质，则其多次测量的结果间的一致性程度就叫信度，有时也叫测量的可靠性。

一般来说，一个好的测量必须具有较高的信度，也即是说，一个好的测量工具，只要遵守操作规则，其结果就不应随工具的使用者或使用时间等方面的变化而发生较大变化。例如，标准的钢尺是测量长度的一种好的工具，只要操作方法得当，无论何时，也无论何人去测量同一张桌子的高度，其结果应是基本一致的。这说明其信度较高。不过，如果所用的是一种具有较大弹性的皮尺，则不同的人或同一个人在不同的时候

去测量同一张桌子的高度，其结果必然会有较大的差异。这说明这种测量的信度不高。

当然，心理测量要比物理测量复杂些，我们不太可能用同一种量表去反复测量一个人的同一种心理特质。例如，某一数学测验就不能反复使用在同一批人身上，否则，测验结果必然会越测越好。因此，信度的定义还应寻求更实际一些的办法，以下就是另外3种等价的信度定义。

定义1：信度乃是一个被测团体的真分数的变异数与实得分数的变异数之比。即：

$$r_{xx} = s_T^2 / s_x^2 \tag{4.1}$$

式中 r_{xx} 代表测量的信度，s_T^2 代表真分数变异，s_x^2 代表总变异数，即实得分数的变异。

定义2：信度乃是一个被试团体的真分数与实得分数的相关系数的平方。即：

$$r_{xx} = \rho_{Tx}^2 \tag{4.2}$$

定义3：信度乃是一个测验X（A卷）与它的任意一个“平行测验”X′（B卷）的相关系数。即：

$$r_{xx} = \rho_{xx'} \tag{4.3}$$

在上述3个定义中，信度是就一批人的数据而言的，并不是用同一种工具反复测量同一个人（定义3除外）。这样一来，定义的操作性程度提高了。不过，真分数是我们不知道的值，是测量的测查对象，因此，定义1和定义2仍只具有理论意义，只有定义3才具有实际意义。

此外，描述测量一致性程度的指标还可以用信度指数（ρ_{xT}），它实际上是信度系数的平方根。

二、信度的作用

信度是衡量一个量表质量高低的重要指标之一，信度不合要求的量表是不能使用的，人们在编制和使用量表时都特别重视测量的信度。具体地说，信度的作用表现在以下几个方面。

1. 信度是测量过程中所存在的随机误差大小的反映

如果信度很低，测量的随机误差就很大，测量的结果就会与真分数发生较大偏差。而且，这种偏差完全是随机决定的，这就让人无法相信测量的结果。值得指出的是，测量中的系统误差与信度无关。因此系统误差只对测量结果产生恒定的影响，而不会使测量结果上下波动。

2. 信度可以用来解释个人测验分数的意义

从理论上讲，一个人的真分数本来是用同一个测验对他反复施测所得的平均值，其误差则是这些实测值的标准差。然而，这种做法是行不通的。因此，我们可以用一个团体（人数足够多）两次施测的结果来代替对同一个人反复施测，以估计测量误差的变异数。此时，每个人两次测量的分数之差可以构成一个新的分布，这个分布的标准差就是测量的标准误，它是此次测量中误差大小的客观指标，有了这一指标，我们就可以对团体中任何一个人的测验成绩做出恰当的解释（即，能通过区间估计的办法指出测量的精度）。一个测量的标准误可用下式计算：

$$SE = Sx\sqrt{1 - r_{xx}} \qquad (4.4)$$

（式中 SE 为测量的标准误，Sx 为实得分标准差，r_{xx}是测量的

信度。)

3. 信度可以帮助进行不同测验分数的比较

通常，来自不同的测验的原始分数是不能直接进行比较的，而必须转化成标准分数再进行比较。具体办法是采用“差异的标准误”来进行差异的显著性检验，其公式为：

$$SE_d = S\sqrt{2 - r_{xx} - r_{yy}} \qquad (4.5)$$

(式中，S 为相同尺度（如 T 分数的 S = 10）的标准分数的标准差，r_{xx}和 r_{yy}分别是两个测验的信度系数。)

值得指出的是：①一个测验可以有多个信度估计值，因而其误差估计值也会有多个，在实际工作中要注意选择。②本理论假定同一个团体中所有人的测量误差都相同的，但实际上水平高的人与水平低的人在做测量时会有不 同的随机误差。③测量的结果不能僵硬地看成是一个点，而应看成是一个以该点为中心，以 SE 的某个倍数为半径上下波动的一个范围（区间估计)。

第二节　信度的估计方法

信度是反映测量中随机误差大小的指标。由于造成测量的随机误差的方式或来源多种多样，所以信度的估计方法也多种多样。下面所介绍的信度估计方法是分别测量信度的某一方面的，使用时要特别注意它的含义及适用范围。

一、重测信度

1. 含义和计算

重测信度（test - retest reliability）指的是用同一个量表对同一组被试施测两次所得结果的一致性程度，其大小等于同一组被试在两次测验上所得分数的皮尔逊积差相关系数（详见有关统计书）：

$$r_{xx} = [\sum(x-\bar{x})(y-\bar{y})] / \sqrt{\sum(x-\bar{x})^2 \cdot \sum(y-\bar{y})^2} \quad (4.6)$$

（式中 x 及 $\bar{x}$ 是第一次测量的实得分及实得分的平均值，y 及 $\bar{y}$ 是第二次测量的实得分及实得分的平均值，r_{xx}是重测信度）

当信度值较大时，说明前、后两次测量的结果比较一致，测量工具比较稳定，被试的心理特质受被试状态和环境变化的影响较小。用这种测量结果来预测人在短期内的情况是比较好的，因为该结果具有较好的跨时间上的稳定性。

2. 使用的前提条件

重测信度的特点是用同一工具对同一批人测了两次，因此，它只能在允许重测的情况下进行计算。具体地说，它必须满足 3 个条件：①所测量的心理特性必须是稳定的。例如，一人成人的性格特点一般是稳定的，所以许多人格测验常使用重测信度。但是，刚入学儿童的识字量是极不稳定的，只要两次施测的间隔时间稍长，儿童的识字量就会有很大变化。因此，重测信度不能用于这种情况，因为测量结果的不一致很可能是被试水平的变化所致，而不能说明测量工具是否稳定。②遗忘和练习的效果基本上相互抵消。在做第一次测验时，被试可能

会获得某种技巧，但只要间隔的时间适度，这种练习效果会基本上被遗忘掉的。至于两次测验的间隔时间，可以是几分钟，几小时，也可以是几个月甚至是几年，这要根据问题的性质和测量目的而定。通常，智力测验的间隔时间一般在6个月左可。③在两次施测的间隔时期内，被试在所要测查的心理特质方面没有获得更多的学习和训练。这一点，也实际上是要保证被试具有稳定的心理特质。

值得注意的是，同样一个量表，随着第二次测量的时间不同，它可以有不同的重测信度。因此，在报告重测信度时，应说明两次施测的间隔，以及在此期间内被试的有关经历。例如，在中国修订《韦氏儿童智力量表手册（C—WISC)》中，就曾对重测信度的计算报告了被试情况（6~16岁城市儿童151名，农村儿童74名且各年龄儿童分配较均匀），并报告了两次测验的间隔时间（2~7周）以及两次的相关系数（城市：0.59~0.86，农村：0.59~0.81）等。

二、复本信度

1. 含义与计算

复本信度（Alternate－form reliability）指的是两个平行的测验测量同一批被试所得结果的一致性程度，其大小等于同一批被试在两个复本测验上所得分数的皮尔逊积差相关系数。

不过，两个复本测验实施的时间不同，复本信度所表达的含义略有不同。如果两个复本测验是同时连续施测的，则称这种复本信度为等值性系数。等值性系数的大小主要反映着两个复本测验的题目差别所带来的变异情况。如果两个复本测验是

相距一段时间分两次施测的，则称这种复本信度为稳定性与等值性系数。此时，两个题目间的差别、两次施测时的情境、被试特质水平等方面的差别都会成为测验结果不一致的重要原因。与其他信度系数相比，此种复本信度最小，也即是说，稳定性与等值性系数是对信度的最严格的检验，其值最低。（在实际工作中，为抵消施测的顺序效应，一般可以随机地选出一半被试先做 A 卷后做 B 卷，另一半被试先做 B 卷后做 A 卷。）

2. 使用前提条件

计算复本信度的条件之一是首先要构造出两份或两份以上真正平行的测验（即 A、B 卷）。什么样的测验才称得上真正平行的呢？这就是：复本测验之间必须在题目内容、数量、形式、难度、区分度、指导语、时限以及所用的例题、公式和测验等其他方面都相同或相似。换句话说，平行测验就是那种用不同的题目测量同样的内容而且其测验结果的平均值和标准差都相同的两个测验。显然，严格的平行测验是很难构造出来的。

计算复本信度的条件之二便是被试要有条件接受两个测验。这种条件主要取决于时间、经费等几个方面。

另外，在使用复本信度时，虽然能克服重测信度的一些缺点，但被试在做第二测验时仍会受到练习和记忆等因素的影响、一些解题的策略等技能技巧也会产生迁移效应。对于稳定性与等值性系数，在报告结果时，也应报告两次施测的间隔，以及在此间隔内被试的有关经历。

三、分半信度

1. 含义及计算

分半信度（split - half reliability）指的是将一个测验分成对等的两半后，所有被试在这两半上所得分数的一致性程度。

分半信度可以和等值性系数一样解释，即可以把对等的两半测验看成是在最短时距内施测的两个平行测验。此外，由于分半信度描述的是两半题目间的一致性，所以它有时也被称作内部一致性系数。

分半信度的计算方法和等值复本信度的计算方法类似，只不过被试在两半测验上得分的相关系数只是半个测验的信度，还必须用斯皮尔曼—布朗公式加以校正：

$$r_{xx} = 2r_{hh} / (1 + r_{hh}) \qquad (4.7)$$

式中 r_{hh}为两半分数间的相关系数，r_{xx}为整个测验的信度值。

不过，斯—布公式只有在两半测验分数的变异数（S_a^2 和 S_b^2）相等时才能使用。否则，我们就应选择下述两个等价的公式之一：

（1）弗朗那根（Flanagan）公式：

$$r_{xx} = 2 [1 - (S_a^2 + S_b^2) / S_x^2] \qquad (4.8)$$

式中 S_a^2 和 S_b^2 分别表示所有被试在两半测验上得分的变异数、S_x^2 表示全体被试在整个测验上的总得分的变异数。

（2）卢仑（Rulon）公式：

$$r_{xx} = 1 - S_d^2 / S_x^2 \qquad (4.9)$$

式中 S_d^2 表示同一组被试在两半测验上得分之差的变异数，其

他符号的含义与（4.8）中含义相同。

2. 使用的前提条件及范围

分半信度通常是在只能施测一次或没有复本的情况下使用。而且，在使用斯皮尔曼—布朗公式时要求全体被试在两半测验上得分的变异数要相等。当一个测验无法分成对等的两半时，分半信度不宜使用。

此外，由于将一个测验分成两半的方法很多（如：按题号的奇偶性分半、或按题目的难度分半、或按题目的内容分半等等），所以，同一个测验通常会有多个分半信度值。

四、同质性信度

1. 含义

同质性信度（homogeneity reliability）也叫内部一致性系数，它是指测验内部所有题目间的一致性程度。这里，题目间的一致性含有两层意思，其一是指所有题目都测的是同一种心理特质，其二是指所有题目得分之间都具有较高的正相关。一句话，同质性信度就是一个测验所测内容或特质的相同程度。

当一个测验具有较高的同质性信度时，说明测验主要测的是某一单个心理特质，实测结果就是该特质水平的反映。如果一个测验同质性信度不高，则说明测验结果可能是几种心理特质的综合反映，这时，测验结果不好解释。一种好的办法是把一个异质的测验分解成多个具有同质性的分测验，再根据被试在分测验上的得分分别作出解释。

值得注意的是，一些表面上看起来是测量同一种心理特质的题目，如果其题目间不具有较高的正相关，则不能认为它们

具有同质性。这即是说，测量单一特性是同质性高的必要条件，而非充分条件。反过来，同质性高才是测验测得单一特质的充分条件。我们讨论同质性信度的目的就在于判断一个测验是否测到单一特质，以及估计所测到特质的一致性程度。

2. 计算及适用范围

内部一致性系数的一种粗略估计方法是求测验的分半信度。但因分半方法多种多样，所得结果不太稳定，因此有人建议：计算出所有可能的分半信度，并用其平均值来做为内部一致性的估计值。然而，这种办法太麻烦了，因为所有可能的分半信度的个数简直是个天文数字，计算机都拿它头痛。于是，人们又提出了如下公式：

$$r_{xx} = K\bar{r}_{ij} / [1 + (K-1)\bar{r}_{ij}] \quad (4.10)$$

其中，K 为一个测验的题目个数，$\bar{r}_{ij}$为所有题目间相关系数的平均值。

这一公式实际也是不方便的，因为所有题目间都求相关会比较麻烦。不过，由此却导出了十分方便的库—理信度系数和克龙巴赫 α 系数，现列于如下：

(1) KR_{20}公式：

$$r_{xx} = [K/(K-1)][1 - (\sum p_i q_i)/S_x^2] \quad (4.11)$$

其中，K 是题目数，p_i 为答对第 i 题的人数的比例，q_i 为答错第 i 题的人数的比例，S_x^2 为测验总分的变异。此公式是由库德（G.F.Kuder）和理查德逊（M.W.Richardson）于 1937 年提出的，仅适用于（0、1）记分的测验。

(2) KR_{21}公式：

$$r_{xx} \doteq [K/(K-1)][1 - (K\bar{p}\bar{q})/S_x^2] \quad (4.12)$$

其中，各指标含义与 KR_{20}相同，只是 $\bar{p}$ 与 $\bar{q}$ 分别表示题目的平均通过率和失败率。此公式只有当所有题目的难度接近时才

适用。

（3）克龙巴赫 α 系数：

$$\alpha=[K/(K-1)][1-(\sum S^2_i)/S_x^2] \quad (4.13)$$

其中，S^2_i表示所有被试在第 i 题上的分数变异，其余指标的含义与 KR_{20}相同。此公式是由克龙巴赫（Cronbach）提出的，它不要求测验题目仅是（0、1）记分，可以处理任何测验的内部一致性系数的计算问题。实际上，KR_{20}和 KR_{21}只是 α 的特例，因为在（0、1）记分时有$\sum S^2_i=\sum p_iq_i$。此外，α 值还是所有可能的分半信度的平均值，它只是测量信度的下界的一个估计值。即，α 值大，必有测量信度高；但 α 值小时，却不能断定测量信度不高。

α 值的计算一般按下述步骤进行：①按一定要求抽取 n 个被试的试卷，首先计算出这几个人测验总分的方差 S^2_x。②这几个人在每一题上都会有一个得分，分别求出这几个人在每道题上得分的方差 S^2_i（i=1，2，…，K），并求$\sum\limits_{i=1}^{K}S_i^2$的值。③按公式（4.13）求出 α 值。

例如，某态度量表共 7 题，100 个被试在各题上得分的方差分别是 0.81，0.82，0.79，0.83，0.85，0.76，0.77，测验总分的方差为 14.00，则此测量的 α 信度为：

$$\alpha=\frac{K}{K-1}\left(1-\frac{\sum S^2_i}{S^2x}\right)=\frac{7}{7-1}$$
$$\left(1-\frac{0.81+0.82+0.79+0.83+0.85+0.76+0.77}{14.00}\right)$$
$$=0.70$$

（4）荷伊特信度：

1941 年荷伊特（C·Hoyt）提出用方差分量比描写测验内部一致性的方法：

设有 n 名被试参加一有 K 个项目的测试，测验分数的总

变异可分解为被试间变异 $SS_{人}$，项目间变异 $SS_{题}$ 和人与试题交互作用 $SS_{人x题}$三部分。荷伊特认为可用 $MS_{人}$ 作为被试方差估计值，用 $MS_{人x题}$作为误差方差估计值，并可用下式作为测验信度的估计值：

$$r_{xx} = 1 - \frac{MS_{人x题}}{MS_{人}} \tag{4.14}$$

五、评分者信度

1. 含义及计算

评分者信度（scorer reliability）指的是多个评分者给同一批人的答卷进行评分的一致性程度。在心理与教育测量工作中，客观题的评分很少出现误差（如机器阅卷），但主观题的评分常常会造成误差。因此，提高评分者信度也是心理与教育测量的重要任务之一。

当评分者人数为 2 时，评分者信度等于两个评分者给同一批被试的答卷所给分数的相关系数（积差相关或等级相关）。

当评分者人数多于两个时，评分者信度可用肯德尔和谐系数进行估计。其公式为：

$$W = 12\left[\sum R_i^2 - (\sum R_i)^2/N\right] / \left[K^2 (N^3 - N)\right] \tag{4.15}$$

其中，K 是评分者人数，N 是被评的对象数（通常是考生数，每个考生一份试卷），R_i 是第 i 个被评对象（考卷）被评的水平等级之和。

当评分者（K）为 3～20 人，被评对象（考卷 N）为 3～7 个时，信度是否合要求可直接查 W 表检验。当实际计算的 W 值大于表中的相应值时，说明评分所得信度较高。

若被评对象多于 7 个，则可计算 X^2 值，作 X^2 检验（$X^2 = K(N-1)W$，$df = N-1$）。

若评分中有相同等级出现，则要使用以下公式求 W 值：

$$W = 12\left[\sum R_i^2 - (\sum R_i)^2/N\right] / \left[K^2(N^2-N) - K\sum\sum(n^3-n)/12\right] \quad (4.16)$$

其中，n 为相同等级的个数，其他指标与（4.15）含义相同。

2. 举例

设有 A、B、C 三位教师给 6 篇作文评分，结果如下，试求评分者信度。

	1	2	3	4	5	6
A	25	30	27	20	28	32
B	22	26	21	20	25	30
C	15	20	18	14	21	22

解：(1) 将每一评分者给 6 篇文章所评分数转化成得分等级（最高分为 1、次为 2，……），并求出每一篇文章所得等级之和 R_i

	1	2	3	4	5	6
A	5	2	4	6	3	1
B	4	2	5	6	3	1
C	5	3	4	6	2	1
R_i	14	7	13	18	8	3

(2) 由上可得：

$\sum R_i = 14 + 7 + 13 + 18 + 8 + 3 = 63$

$\sum R_i^2 = 14^2 + 7^2 + 13^2 + 18^2 + 8^2 + 3^2 = 811$

又由题意知 $K = 3$，$N = 6$

(3) 将 K、N、$\sum R_i$ $\sum R_i^2$ 代入公式（4.15）有：

$$W=\frac{12\ (811-63^2/6)}{3^2\times\ (6^3-6)}\approx 0.95$$

第三节　提高测量信度的方法

一、影响测量信度的主要因素

测量信度是测量过程中随机误差大小的反映。随机误差大，信度就低，随机误差小，信度就高。因此，在测量过程中凡是能引起测量的随机误差的因素——被试、主试、测试内容、施测情境等都会影响测量信度，现具体讨论如下：

（一）被试方面

就单个被试而言，被试的身心健康状况、应试动机、注意力、耐心、求胜心、作答态度等会影响测量误差，因为这些因素往往会影响被试心理特质水平的稳定性。

就被试团体而言，整个团体内部水平的离散程度以及团体的平均水平都会影响测量信度。这是因为，我们所计算的信息估计值大都是以相关为基础的，而相关系数的大小往往取决于全体被试得分的分布情况。当被试团体异质（即团体内水平彼此差异大）时，全体被试的总分分布必然较广，以相关为基础计算出来的信度值必然会大。这就很有可能高估实际的信度值。当团体内部水平相差不大（同质）时，其得分分布必定会较窄，以相关为基础计算出来的信度值必然会小。这时又有可

能低估真正的信度值。此外，若团体的平均水平太高（大家都得高分）或太低（大家都得低分），同样会使测验总分的分布变窄，低估测量的真正信度。

（二）主试者方面

就施测者而言，若他不按指导手册中的规定施测，或故意制造紧张气氛，或给考生一定的暗示、协助等，则测量信度会大大降低。

就阅卷评分者而言，若评分标准掌握不一，或前紧后松，甚至是随心所欲，则也会降低测量信度。

（三）施测情境方面

在实施测验时，考场是否安静，光线和通风情况是否良好，所需设备是否齐备，桌面是否合乎要求，空间阔窄是否恰当等等都可能影响到测量的信度。

（四）测量工具方面

以测验为代表的心理与教育测量工具是否性能稳定是测量工作成败的关键。因此，弄清楚影响测量工具稳定性的因素是十分重要的。一般地，试题的取样、试题之间的同质性程度、试题的难度等是影响测验稳定性的主要因素。

如果一个测验的试题取样不当（或题目数太少、或考察的方面不全面），则难以测查到被试心理特质的全面情况。若是被试采取押题方式应考，则所得结果的随机性更大。既然心理特质被考察的方面是随机的、测查的结果也是随机的，那么，这种测量的信度就必然不会高。相反，当一份测验中的同质性的题目数量增多之后，同一心理特质被考察到的次数就会增多，被试的成绩也就越能被有效地拉开，整个团体的测验分数

分布就会更广，从而提高测量的信度。这种效果可用斯皮尔曼—布朗公式计算：

$$r_{xx} = Kr_{xx} / [1 + (K-1) r_{xx}] \quad (4.17)$$

其中，K 为改变后的测验长度与原来长度之比，r_{xx}为原测验的信度，r_{xx}为测验长度增加为 K 倍后的测验的信度。

如果一个测验内部的试题之间彼此异质（即测查的是不同的心理特质），则无法使测量的内部一致性系数提高。

如果测验的题目太难，则会引起被试的随机猜答，并使得大部分人的得分偏低，整个分数的分布变窄，从而影响测量的信度。如果题目太容易，则大部分被试都将获得高分，同样会使测验分数的分布变窄，影响测量的信度。

（五）两次施测的间隔时间

在计算重测信度和稳定性与等值性系数（复本信度之一）时，两次测验相隔时间越短，其信度值越大；间隔时间越长，其他因素带来影响的机会就多，因而其信度值就可能越小。

二、提高测量信度的常用方法

（一）适当增加测验的长度

由于项目数量太少会降低测量的信度，所以，提高测量信度的一个常用方法是增加一些与原测验中项目具有较好的同质性的项目，增大测验长度。

这里有两点必须注意：①新增项目必须与试卷中原有项目同质。②新增项目的数量必须适度。事实上，增加测验长度的

效果遵循报酬递减规律。即：测验过长有可能引起被试的疲劳和反感，降低测量信度。若已知测验的现有信度，而且知道所要求的信度标准，则可以代入公式（4.17）之中求出 K 值，得到一个恰当的增加数目。

（二）使测验中所有试题的难度接近正态分布，并控制在中等水平。

当测验中所有试题的难度接近正态分布并控制在中等水平时，被试团体的得分分布也会接近正态分布，且标准差会较大，以相关为基础的信度值必然也会增大。

（三）努力提高测验试题的区分度

区分度是测验题目的质量指标，本书稍后将有专论。一份测验所有试题区分度高低直接影响测验的信度。努力提高测验中所有试题的区分度，可望获取较高的测验信度。

（四）选取恰当的被试团体，提高测验在各同质性较强的亚团体上的信度。

由于被试团体的平均水平和内部差异情况均会影响测量信度，所以在检验测量的信度时，一定要根据测验的使用目的来选择被试。即：在编制和使用测验时，一定要弄清楚常模团体的年龄、性别、文化程度、职业、爱好等等因素。一个特别异质的团体上获得的信度值并不等于其中某些较同质的亚团体的信度值。只有各亚团体上信度值都合乎要求的测验才具有广泛的应用。

（五）主试者严格执行实测规程，评分者严格按标准给分，实测场地按测验手册的要求进行布置，减少无关因素的干扰。

三、几点说明

（1）提高测量信度的方法还有很多，以上只是其中的几种常用方法。

（2）本章所讨论的各种信度计算方法仅适用于常模参照性测验。

（3）目标参照性测验的信度问题必须以测量的概化理论（Generalizability Theory 简称 GT）为基础才能进行较好的处理，所以本章未对此进行讨论。此外，速度测验的信度问题也未作讨论，但这并不意味着这些内容不重要。

（4）关于测量的信度要达到多高才被认为可靠？这是一个比较复杂的问题，我们在此给出几个一般性标准供读者参考：标准化能力或学绩测验信度应在 0.90 以上，人格测验的信度应在 0.80 以上，教师自编学绩测验的信度能达到 0.60 以上，就应认为是较高信度的测验了。

练习与思考

1. 指出各种信度系数所对应的误差来源。

2. 已知 16 人参加一次测验后在奇数题和偶数题上的得分情况，试用两种以上方法估计测量信度。

被试	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
奇数题	32	40	42	28	35	30	41	28	32	34	26	34	36	25	40	41
偶数题	31	39	45	30	40	29	39	30	32	30	30	40	36	26	40	42

3. 已知某态度量表有 6 道题，被试在各题上得分的方差分别是 0.80，0.81，0.79，0.78，0.80，0.82，测验总分的方差为 16.00，求 α 值。

4. 怎样提高测量信度？

5*. 试证信度三个定义的等价性。

6*. 不同能力水平的人在接受同一测验时，为什么会有不同的测量误差？

第五章　测量效度

本章提要：

- 效度的概念及其与信度的关系
- 效度的种类及效度获取的方法
- 提高测量效度的方法

在测量活动中，测量者对所使用的测量工具非常信任，他会采取复测行为以判断测量有无误差；如果测量者对所使用的测量工具发生怀疑，那他往往会去找一公认非常准确的测量工具对先前的测值进行检验。这种在原测量工具之外寻求新的证据来肯定或否定某一测量工具准确性的做法就是在研究测量的效度问题。心理测量是一种间接测量，心理测量更重视测量的效度研究。

第一节 效度概述

一、什么是效度

效度（Validity）是指一个测验或量表实际能测出其所要测的心理特质的程度。例如，一个小学生数学测验的成绩若同时受到其数学和语文能力的影响（如，有的人看不懂题意等），则认为实际测到其所要测的特质（数学能力）的程度不高，因而它是个效度不高的数学测验。

关于效度的概念，我们要特别注意以下几点：

(1) 效度是一个相对的概念。这种相对性表现在两个方面：

①效度是相对于一定的测量目的而言的。因为效度是指实测结果与所要测查的特质之间的吻合一致性程度，因此，一个测验或量表是否有效主要是看它是否达到了测量目的。测量某一特质有效的量表，若用它来测量另一种特质，则必然会无效

或效度极低。例如，测量身高很有效的钢尺若用它来测量体重则必定是无效的。又如，人的测量智力很有效的量表若是用来测量性格则必定是效度不高的。②心理特质是较隐蔽的特性，只能通过他的行为表现来进行推测，因此，心理测量不可能达到百分之百的准确，而只能达到某种程度上的准确。不过，由于任何一个量表的编制都有其目的，所以在正常情况下，一个量表的效度也不会为零。例如，一个数学测验，无论其文字表达如何艰深，它总能测到一定的数学能力，即总会有一定的效度，而不会效度为零。

（2）效度是测量的随机误差和系统误差的综合反映。

当一个测验随机误差较大时，实测结果当然会偏离真值，造成结果的不准确。如果测量中还存在系统误差，则系统误差也会加大测量误差。无论出现哪种情况，也无论是否两种误差都存在，只要出现测量误差，测量的效度必受影响。

（3）判断一个测量是否有效要从多方面收集证据。

表面看来，测量的效度就是实际测量的结果与我们所要测量的心理特性的吻合一致性程度，获取效度的办法也就是拿实测结果与心理特性来比较。然而，心理特性是我们要测的东西，是未知的，通常也是比较抽象和隐蔽的。因此，不能把它直接拿来与结果比较，而必须先从多种角度把这种特性描述清楚。由于描述心理特性的角度可以是理论上的，也可以是实践上的，途径很多，因此，获取测量效度的途径也是多样的。例如，智力测验是否测得了人的智力，我们就可以从理论上做逻辑分析，也可以从他在工作、学习中的实际表现等许多方面加以证实。

在前一章曾讲到，一组测验分数的总变异包括3部分：真实的（稳定的）、与测量目的有关的变异；真实的、但出自无关来源的变异；随机误差的变异。

在测量理论中，效度被定义为：在一列测量中，与测量目的有关的真实变异数（由所要测量的变因引起的有效变异）与总变异数（实得变异数）的比率，即：

$$效度 = S_V^2/S_X^2 = r_{xy}^2 \quad (5.1)$$

这里 r_{xy} 代表测量的效度系数，S_V^2 代表有效变异数，S_X^2 代表总变异数。

一个测验的效度表明，在一组测验分数中，有多大比例的变异是由测验所要测量的变因引起的。和信度一样，效度也是指的一列测量的特性，也是一个构想的概念。

二、效度与信度的关系

根据公式 $S_X^2 = S_V^2 + S_I^2 + S_E^2$，可以得到信度与效度的关系如下：

（一）信度高是效度高的必要而非充分的条件

当随机误差的变异数（S_E^2）减小时，真实分数的变异数（S_T^2）增加，测验信度（S_T^2/S_X^2）随之提高。信度的提高只给有效变异数（S_V^2 的增加提供了可能）至于是否能提高效度，还要看系统误差变异数（S_I^2）的大小。可见，信度高不一定效度高。但一个测验要想效度高，真分数的变异数必须占较大的比重，即测验的信度必须高。

信度和效度的这种关系，从日常经验中也可以看。一个测量工具具有一定的信度，但对于某一个目的并不一定是有效的；而一个测量工具如果对于某一个目的是有效的，那么它一

定是可信的。譬如，用米尺来量身高是有效的，也是可信的，而用米尺来量体重，虽然多次量得的结果是一致的，即有较高的信度，但它的效度却很低。

（二）测验的效度受它的信度制约

根据效度和信度的定义（$r_{xy}^2 = S_V^2/S_X^2$，$r_{XX} = S_T^2/S_X^2$）以及公式（$S_T^2 = S_V^2 + S_I^2$）可得到：

$$r_{xy}^2 = (S_T^2 - S_I^2)/S_X^2 = r_{XX} - S_I^2/S_X^2$$

$$\because \quad S_I^2 > 0$$

$$\therefore \quad r_{xy}^2 < r_{xx}$$

这就是说，一个测验的效度总是受它的信度所制约。

第二节　效度的估计

由于测量效度是就测量结果达到测量目的的程度而言的，所以测量效度的估计在很大程度上取决于人们对测量目的的解释。目前，比较常见的解释角度主要有3种，一是用测量的内容来说明目的；二是用心理学上某种理论结构来说明目的；三是用工作实效来说明目的。于是便有了内容效度、结构效度和实证效度之说。当然，这种分类是相对的，一个测验也许需要同时考察它在这3个方面的效度。有些专家甚至认为，效度估计就是多方寻找证据来证明一个测验的有效性程度的过程。本节将着重介绍内容效度、结构效度和实证效度的含义与估计方法。

一、内容效度

1. 内容效度的含义及应用范围

内容效度（Content Validity）是指一个测验实际测到的内容与所要测量的内容之间的吻合程度。估计一个测验的内容效度就是去确定该测验在多大程度上代表了所要测量的行为领域。这里，所要测量的内容或行为领域是依据测量目的而定的，它通常包括欲测的知识范围，以及该范围内各知识点所要求掌握的程度两个方面。首先，在判断一个高中物理试卷是否有较高的内容效度时，我们必须首先分析考题是否有效地覆盖了中学物理所包括的力学、电学、光学、热学以及原子物理 5 个方面。内容效度高的物理测验应当是由这 5 个方面最有代表性的试题样本组成的。其次，我们还必须分析题目的难度等指标是否较好地反映了考试大纲中对这 5 个方面能力水平的要求，等等。

显然，内容效度主要应用于成就测验，因为成就测验主要是测量被试掌握某种技能或学习某门课程所达到的程度的。在这种测验中，题目取样的代表性问题是内容效度的主要考察方面。内容效度高，则可以把被试在该测验上的分数推论到他在相应的知识总体上去，说他在某个方面水平处在一个什么样的位置。反之，内容效度低，则这种推论将是无效的。

内容效度也适合于某些用于选拔和分类的职业测验。这种测验所要测的内容就是实际工作所需的知识和技能，编制这种测验应首先对实际工作做较细的分析，否则，题目取样的代表性就难以令人满意。

应该指出的是，内容效度不适合用于能力倾向测验和人格测验。

此外，在使用内容效度时，要避免与表面效度（surface validity）相混淆。其实，表面效度不能算是一种效度，它不反映测验实际测量的东西。它是外行人对某个测验从表面上看好像是测某种心理特质的一种现象。当外行人认为某个测验能有效地测得某种心理特质时，该测验就被认为有较高的表面效度。一般来说，最佳行为测验往往表面效度高，其他测验则希望表面效度低。

2. 内容效度的确定方法

内容效度的确定方法主要是逻辑分析法，其工作思路是请有关专家对测验题目与原定内容范围的吻合程度作出判断。其具体步骤是：

（1）明确欲测内容的范围，包括知识范围和能力要求两个方面。这种范围的确定必须具体、详细，并要根据一定目的规定好各纲目的比例。

（2）确定每个题目所测的内容，并与测验编制者所列的双向细目表（考试蓝图）对照，逐题比较自己的分类与制卷者的分类，并做记录。

（3）制定评定量表，考察题目对所定义的内容范围的覆盖率、判断题目难度与能力要求之间的差异，还要考察各种题目数量和分数的比例以及题目形式对内容的适当性等等，对整个测验的有效性作出总的评价。

此外，克龙巴赫（Cronbach）还提出过内容效度的统计分析方法。其具体方法是：从同一个教学内容总体中抽取两套独立的平行测验，用这两个测验来测同一批被试，求其相关。若相关低，则两个测验中至少有一个缺乏内容效度；若相关高，则测验可能有较高的内容效度（除非两个测验取样偏向同一个

方面)。

还有一种判断内容效度的方法是再测法。这种方法的操作过程是：在被试学习某种知识之前作一次测验（如学习电学之前考电学知识），在学过该知识后再作同样的测验。这时，若后测成绩显著地优于前测成绩，则说明所测内容正是被试新近所学内容，进而证明该测验对这部分内容而言具有较高的内容效度。

二、结构效度

1. 结构效度的含义、特点与应用范围

结构效度（Construct Validity）是指一个测验实际测到所要测量的理论结构和特质的程度，或者说它是指测验分数能够说明心理学理论的某种结构或特质的程度。这里，构想或结构是指心理学理论所涉及到的抽象而属假设性的概念或特质，如智力、焦虑、外向、动机等等，它们通常用某种操作来定义，并用测验来测量。例如，吉尔福特（J. P. Guilford）认为创造力是发散性思维的外部表现，是人对一定刺激产生大量的、变化的、独创性的反应能力。根据这一理论，他认为创造力测验应重点测量人的思维的流畅性、灵活性和创造性。测验编好后，若有足够的证据来证明它确实可以测到这些特性，则认为它是个结构效度较高的创造力测验。

根据定义，我们可知结构效度的研究具有如下一些特点：

(1) 结构效度的大小首先取决于事先假定的心理特质理论。一旦人们对同一种心理特质有着不同的定义或假设，则会使得关于该特质测验的结构效度的研究结果无法比较。例如，

同样是智力测验，由于当今理论界对智力持有不同的定义，所以，有些智力测验的结构效度的研究结果是不宜进行比较的。

(2) 当实际测量的资料无法证实我们的理论假设时，并不一定就表明该测验结构效度不高，因为还有可能出现理论假设不成立，或者该实验设计不能对该假设作适当的检验等情况。这就使得结构效度的获取更为困难。

(3) 结构效度是通过测量什么、不测量什么的证据累积起来给以确定的，因而不可能有单一的数量指标来描述结构效度。

与内容效度不同，结构效度主要用于智力测验、人格测验等一些心理测验方面。

2. 结构效度的确定方法

总的来说，结构效度的确立一般包括3步：①提出理论假设，并把这一假设分解成一些细小的纲目，以解释被试在测验上的表现。②依据理论框架，推演出有关测验成绩的假设。③用逻辑的和实证的方法来验证假设。例如，韦氏智力测验就是根据这3步来确立结构效度的。韦克斯勒（Wechsler）首先假定“智力是一个人去理解和应付他的周围世界的总的才能”，而不仅仅是推理能力或其他一些具体的技能。然后，他依据这一定义，编制了11个分测验（WAIS－R）或12个分测验（WISC－R），从十几个方面来说明智力，并声明这些个分测验并非是测量不同类型的智力，而是总的智力的各个方面。测验编好以后，许多研究者便从众多角度研究了它的效度。其中，用因素分析方法得出的结论是，该测验实质上测量了三类共同因素，即A因素（言语理解因素）、B因素（知觉组织因素）和C因素（记忆和注意集中因素）。

具体地说，结构效度的估计可以有以下一些方法：

(1) 测验内部寻找证据法。首先，我们可以考察该测验的

内容效度，因为有些测验对所测内容或行为范围的定义或解释类似于理论构想的解释，所以，内容效度高实质上也说明结构效度高。例如，在编制语文能力测验时，许多编制者给内容的定义等同于“语文能力”的解释。其次，我们可以分析被试的答题过程。若有证据表明某一题目的作答除了反映着所要测的特质以外，还反映着其他因素的影响，则说明该题没有较好地体现理论构想，该题的存在会降低结构效度。例如，有些表面上是测人的性格的题目，实质上还涉及到了较多的道德观念，则认为该题会降低性格测验的结构效度。再次，我们足可以通过计算测验的同质性信度的方法来检测结构效度。若有证据表明该测验不同质，则可以断定该测验结构效度不高。当然，测验同质只是结构效度高的必要条件。

(2) 测验之间寻找证据法。首先，我们可以去考察新编测验与某个已知的能有效测量相同特质的旧测验之间的相关。若二者相关较高，则说明新测验有较高的效度。这种方法叫相容效度法。其次，我们也可以去考察新编测验与某个已知的能有效测量不同特质的旧测验间的相关。若二者相关较高，则说明新测验效度不高，因为它也测到了其他心理特质。值得说明的是，二测验间相关不高只是新测验效度较高的必要条件，并不是充分条件。这种方法也叫区分效度法。再次，我们还可以通过因素分析的方法来了解测验的结构效度。其原理是：通过对一组测验进行因素分析，找出影响测验的共同因素。每个测验在共同因素上的负荷量（即测验与各因素的相关）就是测验的因素效度，测验分数总变异中来自有关因素的比例即是该测验结构效度的指标。例如，一些研究者对 WISC－R 和 WISC－CR 作因素分析后，发现公共因子有三个。并且其中的 A 因子的主要负荷测验为词汇、分类、知识和领悟，B 因子的主要负荷测验为图片排列、木块图、填图和图形拼凑，C 因子的主要

负荷测验为算术、数字广度和编码。

(3) 考察测验的实证效度法。如果一个测验有实证效度，则可以拿该测验所预测的效标的性质与种类作为该测验的结构效度指标，至少可以从效标的性质与种类来推论测量的结构效度。这里有两种做法：其一是根据效标把人分成两类，考察其得分的差异。例如，一组被公认为是性格外向的人在测验中得分较高，另一组被公认为是性格内向的人在测验中得分较低，则说明该测验能区分人的内向与外向特征，进而说明该测验在测量人的性格内外向方面有较高的结构效度。其二是根据测验得分把人分成高分组和低分组，考察这两组人在所测特质方面是否确有差异。若两组人在所测特质方面差异显著，则说明该测验有效，具有较高的结构效度。此外，对于一些被认为是较稳定的特质，若在短期内两次施测的结果差异不太大，则说明该测验符合理论构想。

(4) 多种特质——多种方法矩阵法。该方法实质是相容效度和区分效度法的综合运用，其原理是若用多种极不相同的方法测量同一种特质相关很高（用极为相似的方法测量不同特质相关很低），则说明测量效度较高。于是，若有多种特质（如A、B、C）都接受了多种方法（如1、2、3、4）的测查，就可以分别计算出任意两种方法测量同一特质的相关和测量不同特质的相关，以及任意两种特质接受同一方法和不同方法的相关，并以这些相关系数为元素构成一个矩阵，如下图所示：

在上表中，位于主对角线上的数值，是用同样的方法测相同特质所得的相关，是信度指标；在实三角形内的数值，是用同样方法测不同特质所得之相关。此相关若高，则说明方法间共同点较多；在虚线三角形内的数值，是用不同方法测量不同特质所得的相关，它一般较低，是特质与方法间交互影响的反映；在虚线三角形之间的两条对角线上的数值，是用不同方法

测相同特质的相关，它是测验效度的指标。

表 5.1 多种特质——多种方法矩阵

		方法 1			方法 2			方法 3			方法 4		
	特质	A_1	B_1	C_1	A_2	B_2	C_2	A_3	B_3	C_3	A_4	B_4	C_4
方法 1	A_1	.90											
	B_1	.50	.89										
	C_1	.35	.41	.81									
方法 2	A_2	.58	.25	.10	.95								
	B_2	.21	.59	.09	.63	.91							
	C_2	.14	.13	.50	.57	.53	.85						
方法 3	A_3	.55	.20	.13	.69	.32	.30	.93					
	B_3	.11	.60	.19	.20	.68	.29	.50	.96				
	C_3	.15	.20	.70	.21	.19	.67	.53	.51	.92			
方法 4	A_4	.58	.21	.11	.66	.11	.19	.70	.13	.14	.89		
	B_4	.18	.61	.09	.30	.68	.18	.22	.68	.20	.51	.90	
	C_4	.20	.15	.71	.22	.18	.70	.23	.19	.71	.52	.50	.91

三、实证效度

1．实证效度的含义、种类及作用

实证效度是指一个测验对处于特定情境中的个体的行为进行估计的有效性。也就是说，一个测验是否有效，应该以实践的效果来作为检验标准。例如，当我们用机械能力倾向测验测查了一大批机械工人之后，若有证据表明测验高分组的实际工作成绩确实优于低分组的实际工作成绩，则可以认为该测验具有较高的实证效度。又如，在军队选拔汽车驾驶兵时，若用测验选出来的兵在学习驾驶技术，以及日后的驾驶过程中的表现都大大好于以前未用测验随意指派的汽车兵，则表明该测验也具有较高的实证效度。

在这里，被估计的行为是检验测验效度的标准，简称为效标。实证效度主要重视那些与测验独立的效标行为，而不太注重测验内容或结构。实证效度也称效标关联效度。

根据效标资料搜集的时间差异，实证效度可以分成同时效度和预测效度两种。例如前文所说的机械能力倾向测验，其效标资料是与测验分数同时搜集的，所以它是同时效度。前文中所说的汽车兵选拔测验，其效标资料是在测验之后根据实际工作成绩来确定的，所以它叫预测效度。

同时效度主要用于诊断现状，在于用更简单、更省时、更廉价和更有效的测验分数来取代不易搜集的效标资料。预测效度的作用在于预测某个个体将来的行为。无论是同时效度还是预测效度，其目的都是想通过对测验在一个有代表性的样本上，用实证的方法来证明测验有效，于是在今后就可以用简便

的测验去预测类似于样本的其他团体或个体的行为。因此，有人把二种效度都称作预测效度，并把测验称作预测源。

2. 效标

估计测验的实证效度的首要条件是必须具有效标，那什么是效标？效标如何表达？

简单地说，效标就是衡量一个测验是否有效的外在标准，它是独立于测验并可以从实践中直接获得的我们所感兴趣的行为。

不过，我们所感兴趣的行为往往是一个观念上的东西（观念效标），它必须用一个数字或等级来进行表达（效标测量）。例如，大学入学考试的观念效标通常是“大学学习成功”，它的一种常用的效标测量便是大学头两年或一年相关学科的平均成绩。

显然，同一个观念效标可以有多个效标测量（多样性），而且每一种效标行为往往都是由多种特质构成，因此效标测量是件极为复杂的事（复杂性）。又因效标测量有多种多样，所以有些效标测量只可以反映测验在某一特殊方面的有效性程度，即，在一种情况下有效的测量，在另一种情况下未必有效（特殊性和时间性）。这就要求测验的编制者和使用者要特别小心。

一般说来，效标测量要想较好地体现观念效标，那效标测量本身就必须是有效的和可靠的，而且还必须客观、实用。

在心理与教育测量工作中，常用的效标主要有：学业成就、等级评定、临床诊断、专门的训练成绩、实际的工作表现、对团体的区分能力以及其他现成的有效测验。这些效标可以是连续变量，也可以是离散型变量；可以是自然的现成指标，也可以是人为设计的指标；可以是主观判断，也可以是客观测量；可以是自我评定，也可以是他人评定等等。

3．实证效度的确定方法

实证效度的确定方法大体上可以分为以下几个步骤：①明确观念效标。②确定效标测量。③考察测验分数与效标测量的关系。

从效度估计的方法上看，实证效度可以用以下方法进行估计：

（1）相关法。

实证效度的一种常用估计方法是计算测验分数与效标测量的相关系数（积差相关法、等级相关法、二列相关法、四分相关法等等）。例如，张厚粲教授在主持修订瑞文标准推理测验（SPM）时，她报告的同时效度就是北京一所普通中学45名12～15岁学生同时接受SPM和韦氏儿童智力测验得分的积差相关系数，预测效度则是对北京市两所中学69名高三学生先施测SPM，再搜集这批学生3个月后的高考成绩，最后计算SPM成绩与高考语文、数学和总分的积差相关。

（2）区分法。

该方法的思路是，被试接受测验后，让他们工作一段时间，再根据工作成绩（效标测量）的好坏分成两组。这时再回过头来分析这两组被试原先接受测验的分数差异，若这两种人的测验分数差异显著，则说明该测验有较高的效度。

（3）命中率。

当用测验作取舍决策时，决策的正命中率和总命中率是测验有效性的较好指标。其中，总命中率是指根据测验选出的人当中工作合格的人数，以及根据测验淘汰的人当中工作不合格的人数之和与总人数之比。若总命中率高，则说明测验的效度高。这种测验在区别合格与不合格方面是有效的。此外，有些测验只关心被选者中合格者有多少，而不关心被淘汰者中是否有合格者。这时测验的效度应该用测验的正命中率来评价。所

谓正命中率是指用测验选出的人中合格者所占的比例。这个比例越高，测验越有效。

不过，在评价一个测验的效度时，还要注意测验使用的功利率比例，即：使用测验所带来的好处应大大高于使用测验所耗费的时间、精力和经费，还要比较用测验与不用测验的效益之差，若差别不大，则无使用测验之必要。

第三节　提高测量效度的方法

一、影响测量效度的因素

严格地说，凡是与测量目的无关的稳定的和不稳定的变异来源都会影响测量的效度。这就是说，测验本身的构成、受测被试的特点、施测的过程、阅卷评分、分数的转换与解释等一切与测量有关的环节都可能影响测量的效度。现择其主要方面给予说明。

1. 测验的构成

当组成测验的试题样本没有较好地代表欲测内容或结构时，测量的内容效度或结构效度就必然会不高。同时，若题目语义不清、指导语不明、题目太难或太易、题目太少或安排不当等等，都会降低测量效度。一般而言，增加测验的长度可以提高测量信度，进而为提高测量效度提供了可能。于是，一些研究者便得出了测验长度与效度的公式如下：

$$r_{(Kx)y}=\frac{Kr_{xy}}{\sqrt{K(1-r_{xx}+Kr_{xx})}} \qquad (5.2)$$

式中 $r_{(Kx)y}$是测验 x 增长至原来的 K 倍后，新测验与效标（y）的相关（效度系数）；K 为测验增长的倍数；r_{xy}为原测验的效度系数；r_{xx}为原测验的信度系数。

2. 测验的实施过程

一个测验在实施过程中，如不遵从指导语的要求、或出现意外干扰、或评分计分出现差错等等，都会降低测量效度。

3. 接受测验的被试

一般情况下，被试的应试动机、情绪、态度、身体状态等等，都会影响测量信度，造成较大的随机误差，进而影响测量的效度。

就整个被试团体而言，如果缺乏必要的同质性，则很可能会得到不恰当的效度资料。有时候，同样一个测验，对年龄、性别、文化程度、职业等方面不同的被试团体，常常表现出不同的预测能力，即具有不同的测量效度。事实上，被试团体的年龄、性别、文化程度与职业等方面的特征，常常成为干涉变量。我们在考察效度时，要特别注意测验在不同团体上的效果，避免出现测验偏倚（test bias）。

4. 所选效标的性质

由于同一个测验可以有不同的效标，同一个观念效标也可以有不同的效标测量，所以在评价测量效度时，所选效标的性质是很重要的考虑因素。

有的学者指出，智力测验分数与教师对学生等级评定之间的效度系数只要在 0.30～0.50 之间就可以了，因为教师的评价会受到与智力无关的其他因素的影响。与此类似，相同科目的标准化测验成绩与教师评价之间的相关应达到 0.60～0.70，两种不同智力测验或标准化测验之间的相关应达到 0.60～

0.80等等。所有这些不同的要求，主要是因为所用效标的不同而提出来的。

在考虑效标与分数的相关时，有一个因素是必须重视的，即测验分数与效标之间是否符合线性关系的问题。因为皮尔逊积差相关的前提之一是二变量间具有线性关系，否则会得出错误的效度结论。这就要求我们在选用相关系数的计算公式时，注意各公式的使用条件。

5. 测量的信度

前文已经论及，测量信度是测量的随机误差的反映，而任何误差的增加都会降低测量的效度，所以在考察测量效度时，一定要注意测量信度。信度不高的测验不可能具有很高的测量效度。

二、提高测量效度的方法

要想提高测量效度，就必须设法控制随机误差、减小系统误差，同时，还要选择好特别恰当的效标，把效度系数准确地计算出来。具体来说，下述方法能提高测量效度：

(1) 精心编制测验量表，避免出现较大的系统误差。

这就要求题目样本要能较好地代表欲测内容或结构，要避免出现题目偏倚 (item bias)。同时，题目的难易程度、区分度也要恰当，题目的数量也要适中。太难、太易、太多、太少都是有损测量效度的。此外，测验试卷的印制，题目作答的要求，评分计分的标准，题目意思的表述等等，都必须严格检查，避免一切可避免的误差的出现。

(2) 妥善组织测验，控制随机误差。

在测验实施过程中，系统误差一般不太明显，但随机误差却有可能失控。这就要求测验实施者一定要严格按手册指导语进行操作，要尽量减少无关因素的干扰。

(3) 创设标准的应试情境，让每个被试都能发挥正常的水平。

在各种测验中，有些被试往往因种种原因而发挥不出应有水平（比如过分焦虑致使水平失常等），因此，我们应让被试调整好应试心态，让他们从生理上、心理上、学识上等做好应有的准备。否则，焦虑因素和其他无关因素影响过大，必然会降低测量效度，测不到欲测的内容或结构。

(4) 选好正确的效标、定好恰当的效标测量，正确地使用有关公式。

在评价一个测验是否有效时，效标的选择是一个重要方面。假若所选效标不当，或所选效标无法量化，则很难正确地估计出测量的实证效度。如果效标及效标测量都合乎要求，则公式的选择也是影响效度估计的重要方面。

练习与思考

1. 什么是测量效度？它与信度的关系怎样？

2. 什么是内容效度？测验编制者和使用者应分别从哪几个方面来把握内容效度？

3. 什么是结构效度？测验编制者和使用者应分别怎样把握结构效度？

4. 什么是实证效度？它与内容效度和结构效度有何异同？

5. 什么是效标和效标测量？

6. 已知 $r_{xx}=0.31$，$r_{xy}=0.42$，若希望把效度系数提高到0.65和0.70，则测验长度要增加几倍？

7. 复习有关教育与心理统计知识，弄清各种相关系数的计算方法与使用条件。

8*. 假设某学者自编了一套神经类型测验，并且在几十万人中进行了试用，试问该测验是否一定有效？

第六章　测验的项目分析

本章提要：

●测验项目难度的意义，难度指标的计算及项目难度对测验的影响。

●测验项目区分度的意义，区分度的求法、区分度对测验质量的影响。

●测验猜测问题的纷争。

●多重选择题的项目分析方法。

第一节　测验的难度

一、难度的意义

难度是指测验项目的难易程度。一个测验项目，如果大部分被试都能答对，则该项目的难度就小；如果大部分被试都不能答对，则该项目的难度就大。

二、难度的计算

测验的记分方法不同，项目难度的计算方法也有所不同。

（一）二分法记分项目的难度

1. 通过率

如果不考虑被试作答是猜测成功的机遇，二分法记分测验项目的难度通常以通过率来表示，即以答对或通过该项目的人数的百分比来表示：

$$P=\frac{R}{N} \tag{6.1}$$

式中，P 代表项目难度，N 为全体被试数，R 为答对通过

该项目的人数。

例如，在200个学生中，答对某项目的人数为120人，则该项目的难度为 $P=\frac{120}{200}=0.60$。

以通过率表示项目的难度时，通过人数越多，P值越大，其难度越小；通过人数越少，P值越小，难度越大，题目越难。所以有人也称P值为容易度。事实上，这里的P值与我们通常所理解的难度意义正好相反。

2. 极端分组法

当被试人数较多时，则可以先将被试依照测验总分从高到低排列，分成三组，总分最高的27%被试称为高分组（N_H），总分最低的27%被试为低分组（N_L），分别计算高分组和低分组的通过率，然后求项目的难度。

$$P=\frac{P_H+P_L}{2} \tag{6.2}$$

或

$$P=\frac{1}{2}\left(\frac{R_H}{N_H}+\frac{R_L}{N_L}\right) \tag{6.2'}$$

式中 P_H、P_L 分别表示高分组和低分组的通过率；R_H、R_L 表示高分组和低分组通过该项目的人数；N_H、N_L 分别代表高分组和低分组的人数。

例如，在370名被试中，选为高分组和低分组的被试各有100人，其中高分组有70人答对第1题，低分组有40人答对第1题，则第1题的难度为：

$$P=\frac{1}{2}\left(\frac{70}{100}+\frac{40}{100}\right)=\frac{1}{2}(0.70+0.040)=0.55$$

（二）非二分法记分项目的难度

对于论述题，每个项目不只有答对和答错两种可能结果，而是从满分至零分之间有多种可能结果。对这类项目，常常用

下面的公式来计算其难度。

$$P = \frac{\bar{x}}{x_{max}} \tag{6.3}$$

式中 $\bar{x}$ 为被试在某一项目上的平均得分，x_{max} 为该项目的满分。

例如，数学测验的第七题满分的 15 分，该题考生的平均得分为 9.6 分，则该题的难度为：

$$P = \frac{\bar{x}}{x_{max}} = P = \frac{9.6}{15} = 0.64$$

三、测验难度水平的确定

进行难度分析的主要目的是为了筛选项目，项目的难度水平多高合适，取决于测验的目的的项目形式以及测验的性质。

在教育工作或实际工作中，若测验的目的是为了了解被试在某方面知识技能的掌握情况，可以不必过多地考虑难度，只要教育者认为重要的内容就可以选用，甚至那些 100%通过或通过率为 0 的项目都可以采用。例如，在某单元教学之前，要了解学生对所要教学的内容准备情况所作的预备测验，几乎每个项目都将产生很低的通过率，但这些项目不应淘汰，因为它们表明了哪些内容需要学生认真学习并加以掌握。而在教完某部分知识以后，为了检查学生的掌握情况所进行的测验，即使每道项目都有很高的通过率，这些项目仍然是可用的，它们表明学生的掌握程度。

如果测验的目的是用于选拔录用人员，就应该将项目的难度控制在接近录取率左右，即较多地采用那些难度值接近录取

率的项目。例如，要从高中生中选拔15%的人参加全市的数学竞赛，则就应提高项目的难度，使P值接近0.15。

四、难度的等距变换

以项目的通过率来表示项目的难度，虽然计算方便，易于理解，但这类难度指标属于顺序变量，不具有相等的单位，所指出的仅仅是项目的相对难度。例如，3个测题的难度指数分别为0.60、0.70、0.80，我们只能说，第一题最难，第二题次之，第三题最容易。虽然三题难度分别相差10%，但我们并不能说第一题与第二题的难度之差等于第二题与第三题的难度之差。通过率P无法指出难度之间差异大小，可见顺序性这一点，对我们作进一步的难度分析带来了困难，必须设法将它转换成等距量表。

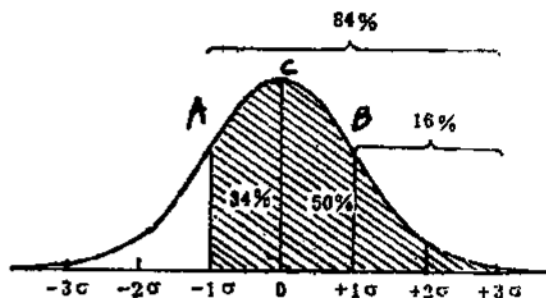


图6.1　正态分布下通过率与工值的关系

当样本容量很大时，测验分数将接近正态分布。此时，我

们可以根据正态分布曲线表，将试题的难度 P 作为正态曲线下的面积，转换成具有相等单位的等距量数，即 Z 分数。我们知道，在正态分布中，平均数之上或之下一个标准差的距离约占全体人数的 34%，因此，如果在一个测验中某项目 A 通过率为 84%（P = 0.84），那么从图（6.1）可以看出，这项目的难度就在平均数以下一个标准差位置，即难度为 -1σ；如果某项目 B 的通过人数只有 16%，（= 0.16）则这个项目的难度为 $+1\sigma$；若某题 C 恰好有 50% 的人通过（p = 0.50），则该题的难度为 0，应用此方法，任何一个与通过率相当的难度值都可以通过查正态分布表得到。显然，较难的项目难度为正值，较易的项目难度为负值。由于标准分数具有相等单位，属于等距量表。所以，用标准分数作为项目难度的指标，为进一步作难度分析带来了极大的方便。

但是，Z 分数有小数点和负值，所以表示难度也有不便之处，通常需要转换成另一种单位的等距量表。其中较为常用的是美国教育测验服务中心采用的难度指标：

$$\triangle = 13 + 4 \cdot Z \qquad (6.4)$$

式中，△表示题目难度，Z 表示由 P 值转换得来的标准分数。

例如，上面所举的例中，其△难度值为：

项目 A：通过率 $P = 0.84$　$Z = -1$　$\triangle = 13 + 4 \times (-1) = 9$

项目 B：通过率 $P = 0.16$　$Z = 1$　$\triangle = 13 + 4 \times 1 = 17$

项目 C：通过率 $P = 0.50$　$Z = 0$　$\triangle = 13 + 4 \times 0 = 13$

根据正态分布表，可以知道，△是以 25 为上限，1 为下限的等距量表，△值愈大，则难度愈高，△值愈小，难度越低。

对一般教师来说，只要计算出 P 值即可。但如果要作更精确的统计分析，则就需要计算出具有等距量表性质的△值。

五、难度对测验的影响

（一）测验难度影响测验分数的分布形态

测验的难度直接依赖于组成测验的项目的难度。通过考察测验分数的分布，可以对测验的难度作出直观分析。

若测验项目的难度普遍较大，被试的得分普遍较低，使得测验分数集中在低分端，其分数分布呈现正偏态；当测验题目的难度普遍较小，被试的得分普遍较高，测验分数集中在高分端，分数分布呈现出负偏态。

测验难度过大或过小，都会造成测验分数偏离正态分布。但是，由于人的多数心理特质是正态分布。而我们目前所采用的统计分析方法（例如前面介绍的难度的等距交换）又大都是以正态分布为前提，所以大多数测验在设计时希望分数呈现正态分布模式。因此，当测验的分数分布为明显偏态时，可通过改变项目难度的比例来加以调整。通常，若被试的取样具有代表性，对于中等难度的测验，其分数分布呈现正态分布。

（二）测验难度影响测验分数的离散程度

过难或过易的测验，会使测验分数相对地集中在低分端或高分端，从而使得分数的全距缩小。1965 年艾伯尔（R.L Ebel）用三套各包含有几个项目的测验进行研究，各套测验的分数分布见图（6.2）[①]，从图形可见，当难度集中在 0.50 附近时，分数的分布范围较广，方差较大（$\sigma=2.6$）；而当难度集中在两端，即不是太难，就是太易时，分数分布范围最小

($\sigma=1.60$)。根据信度公式 $r_{xx}=1-\frac{\sigma E^2}{\sigma_x^2}$ 可知，分数分布范围较广，测验信度较高，反之则信度值较低。可见，项目的难度以集中在 0.50 左右最佳，以集中两极端最差。

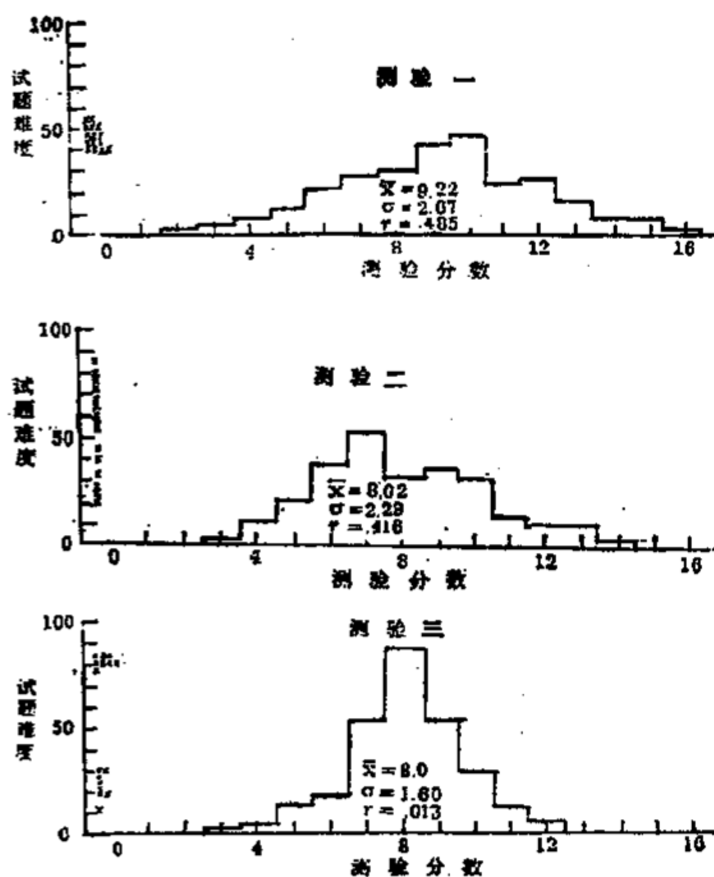


图 6.2　试题难度与测验分数分布的关系

①R. L 艾伯尔《教育测量纲要》淮书青等译，第 274 页（江西师范大学高教研究室）

此外，项目的难度对项目的鉴别能力有一定的联系，这一点将在本章第二节中讨论。

第二节　测验的区分度

一、区分度的意义

区分度是指测验项目对被试心理品质水平差异的区分能力。具有良好区分度的项目，能将不同水平的被试区分开来，也就是说，在该项目上水平高的被试得高分，水平低的被试得低分。反之，区分度低的项目则对不同水平被试不能很好地鉴别，水平高与水平低的被试，所得分数差不多，甚至正好相反。所以测量专家们把试题的区分度称为测验是否具有效度的"指示器"，并作为评价项目质量，筛选项目的主要指标与依据。必须指出：评价测验项目区分度高低依赖于对被试水平的准确测量，通常称作为效标分数。测验项目区分度的效标分数更多的是用测验总分，称作为内部效标。

区分度（D）的取值范围介于 -1.00 至 +1.00 之间。通常 D 为正值，称作积极区分；D 为负值为消极区分；D 为 0 称作无区分作用。具有积极区分作用的项目，其 D 值越大，区分的效果越好。

二、区分度的计算

项目区分度的计算方法很多，各种方法在含义上略有差别。在使用时，我们可以根据测验的目的，以及项目记分和测验总分的两个变量的性质不同，而选择不同的计算方法。当然，有时可以同时用几种计算方法相互验证。

(一) 项目鉴别指数法

这种方法较适合于二分法记分的测验项目。

1. 鉴别指数的计算

当效标成绩是连续变量时，可以从分数分布的两端各选择27%的被试，分别计算出每道题目上的各自的通过率，二者之差就是鉴别度指数（D），即：

$$D = P_H - P_L \tag{6.5}$$

式中 P_H 与 P_L 分别为高分组与低分组在该项目上的通过率。

例如：高分组在某一项目的通过率为 0.75，低分组的通过率为 0.35，则该项目的鉴别指数为 $D = P_H - P_L = 0.75 - 0.35 = 0.40$。当 $D = 1.00$ 时，高分组被试全部通过，低分组被试全部失败。相反，如果低分组的被试全部通过，高分组的被试全部失败，则 $D = -1.00$。如果两组的通过率相等，则 $D = 0$。

D 值是鉴别项目测量有效性的指标，D 值越高，项目越有效。1965 年，美国测验专家 R.L.Ebel 根据长期经验提出用鉴别指数评价题目性能的标准如表 6.1 所示。

表 6.1 题目鉴别指数与评价价标[①]

鉴别指数 D	题目评价
0.40 以上	很好
0.30～0.39	良好、修改会更好
0.20～0.29	尚可、仍需修改
0.19 以下	差、必须淘汰

由于编制测验不容易，一般情况下人们宁愿修改项目，也不愿轻易舍弃项目。当然上述标准也不是绝对的，还必须根据测验的目的、性质、要求来决定项目的取舍。

2. 极端组的划分

在项目难度和鉴别指数分析中多次提到划分高分组与低分组，一般情况下，是根据效标成绩或测验总分将被试排队，取27%的高分端被试组成高分组，另外27%的低分端被试作为低分组，其余46%的被试可以不作分析。有人曾证明（Kelley，1939）当分数分布是正态分布时，这种分配方法很有效，它既可以使两个对比组间的差异尽可能大，又可以使两组人数尽可能多。当效标分数较正态分布平坦时，高低分组各占的比率应略高于27%，约在33%左右。一般情况下，其比率介于25%～33%即可。但如果是标准化测验，习惯上仍采用27%。如果比率太小，如10%，则所选出来的两组过于极端，二者之间的差异非常明显，人为夸大了题目的区分程度；当样本团体人数过少时（n<100），则不宜用27%的规则，甚至可以用50%作为分界点，即把上、下两半被试作为高分组与低组。

使用极端分组法主要是为了计算方便，但是这种方法只利

① R.L.Ebel:《教育测量纲要》，濮书青等译，江西师大高教室印。

用了一部分信息，浪费了很多信息，所以统计结果比用全部资料计算的准确性差一些。当项目与效标之间是直线关系时，这种分析法对结果的准确性来说影响不大。但当项目与效标之间并非直线关系时，使用极端分组法会丧失许多有价值的信息，甚至可能得出错误结论。

（二）相关法

用鉴别指数分析项目区分度虽然易于理解，计算方便，但结果不精确。在大规模的或标准化的测验中，多采用相关法，即以项目分数与效标分数或测验总分的相关作为项目区分度的指标。相关越高，项目区分度越高。

1. 点二列相关

点二列相关适用项目是 0、1 记分（或二分变量），而效标或测验总分是连续变量的数量资料，其计算公式为：

$$r_{pb} = \frac{\bar{x}_p - \bar{x}_q}{S_t}\sqrt{pq} \qquad (6.6)$$

式中：r_{pb}为点二列相关系数；$\bar{x}_p$ 为通过该项目被试的平均效标分数；$\bar{x}_q$ 为未通过该项目被试的平均效标分数；p 为通过该项目被试的人数百分比；q 为未通过该项目被试人数的百分比；S_t 为全体被试的效标分数的标准差。（6.6）式也可以写作 $r_{pb} = \frac{\bar{x}_p - \bar{x}_t}{S_t}\sqrt{\frac{p}{q}}$（6.6′）式中 $\bar{x}_t$ 为全体被试的平均效标分数，余同（6.6）式说明。

例 6.1　15 名被试在某测验第 1 题上的作答情况（通过记 1 分，未通过记 0 分）与效标分数见表 6.2。

表 6.2　15 名被试的效标分数与第一题作答情况

学生序号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
效标分数	65	70	31	49	80	50	35	16	81	69	78	55	77	90	42
第 1 题得分	0	1	0	1	1	0	1	0	0	1	1	0	1	1	0

试计算该测验第 1 题的区分度。

由表 6.2 可以求出：

$$p=\frac{8}{15}=0.5333$$

$$q=1-p=0.4667$$

$$\bar{x}_p=\frac{548}{8}=68.5$$

$$\bar{x}_q=\frac{334}{7}=47.71$$

$$S_t=\sqrt{\frac{\sum x^2}{N}-\left(\frac{\sum x}{N}\right)^2}=\sqrt{\frac{58936}{15}-\left(\frac{882}{15}\right)^2}=21.72$$

$$\bar{x}_t=\frac{\sum x}{N}=\frac{882}{15}=58.8$$

将上述数据代入公式（6.6）或（6.6′）得到：

$$r_{pb}=\frac{\bar{x}_p-\bar{x}_q}{S_t}\sqrt{pq}=\frac{68.5-47.71}{21.72}\sqrt{0.5333\times0.4667}=0.4775$$

或：$r_{pb}=\frac{\bar{x}_p-\bar{x}_t}{S_t}\sqrt{\frac{p}{q}}=\frac{68.5-58.8}{21.72}\sqrt{\frac{0.5333}{0.4667}}=0.4775$

对用点二列相关计算出的数值需进行显著性检验，才能确定其意义。要检验 r_{pb}是否达到显著水平，常用的检验方法有两种：①采用对积差相关系数检验的方法进行检验（可参阅有关统计学教科书）。②用 t 检验的方法比较二分变量对偶的两组连续变量的平均数的差异是否显著，如平均数（$\bar{x}_p$ 与 $\bar{x}_q$）的差异显著，则相关系数也显著。本例若运用第①种方法，可知 r_{pb} 未达到 0.05 的显著性水平，所以该项目的区分度值得怀疑。

2. 二列相关

二列相关适用于连续的测量变量。但其中一个变量因为某种原因被人为分成两类。例如，当一个测验的项目分数是连续的，而效标或测验总分数被分为高低或及格、不及格两个类别时，可以采用二列相关法；当效标或测验总分是连续的，而项目分数被人为分成对、错或通过、未通过两类，也可以采用此方法。其计算公式为：

$$r_b=\frac{\bar{x}_p-\bar{x}_q}{S_t}\cdot\frac{pq}{y}\qquad(6.7)\qquad 或\ r_b=\frac{\bar{x}_p-\bar{x}_q}{S_t}\cdot\frac{p}{y}\qquad(6.7)$$

式中 r_b 为二列相关系数；$\bar{x}_p$、$\bar{x}_q$、$\bar{x}_t$、$\bar{s}_t$、p、q 的意义同点二列相关系数公式（6.6）说明；y 为正态分布下 p 与 q 分割点正态曲线的高度。

例 6.2　仍以前述例 6.1 与表 6.2 的资料，以二列相关法计算区分指数 r_b。

因为：$p=0.5333$，在 p、q 分割点正态曲线高度为 $y=0.3975$（可通过查正态分布表获得）则：

$$r_b=\frac{\bar{x}_p-\bar{x}_q}{S_t}\cdot\frac{pq}{y}=\frac{68.50-47.71}{21.72}\cdot\frac{0.5333\times0.4667}{0.3975}=0.599$$

运用二列相关法求项目区分度时，要求二分变量在人为二分前的测量必须是正态分布，如果样本分布不是正态，总体分布也应该是正态的。对于连续变量的分布，虽不要求是正态但必须是单峰且是对称分布形态。

二列相关系数 r_b 的显著性检验可以用下列公式检验。

$$z=\frac{r_b}{\frac{1}{y}\sqrt{\frac{pq}{N}}}\qquad(6.8)$$

式中，r_b、y、p、q、的意义同前，N 为被试总人数。对上例

$$z=\frac{0.599}{\frac{1}{0.3975}\sqrt{\frac{0.5333\times0.4667}{15}}}=1.85$$

$z=1.85<Z_{0.05}=1.96$，未达到 0.05 的显著性水平，可见计算所得的 r_b 没有达到应有的显著性水平。

3.φ 相关

φ 相关的统计方法适用于两个变量是二点分配的资料，即两个变量都是二分名义变量 。在有些情况下，一些连续变量也可以用此方法计算相关程度。φ 相关不要求变量呈正态分布。所求指标为 φ 系数。

在用 φ 系数作为区分度指标时，要求项目反应与效标变量都是二分状态。一般是根据效标成绩或测验总分的高分组和低分组，通过和未通过某一项目的人数列成的四格表来计算。计算公式为：

$$r_{\varphi}=\frac{ad-bc}{\sqrt{(a+b)\ (c+d)\ (a+c)\ (b+d)}}\qquad(6.9)$$

式中 r_{φ} 为 φ 的相关系数，a、b、c、d 分别为四格表中四项所包含的人次数。

例 6.3　用表 6.2 数据为资料，测验总分以 60 分以上为升级，60 分以下者为留级，就可以归类为下列的 2×2 表。假设以升学情况作为效标，此题对于学生的区分度为多少？

升学情况

	升级	留级	合计
通过	6 (a)	2 (b)	8 (a+b)
未通过	2 (c)	5 (d)	7 (c+d)
	8 (a+c)	7 (b+d)	15N

$$r_{\varphi}=\frac{ad-bc}{\sqrt{(a+b)\ (c+d)\ (a+c)\ (b+d)}}=\frac{6\times5-2\times2}{\sqrt{8\times7\times8\times7}}=\frac{26}{56}=0.4643$$

φ相关的显著性检验可以用 r_{φ} 与 X^2 的关系式求出，并作 X^2 检验。

$$X^2=N\cdot r_{\varphi}{}^2 \tag{6.10}$$

对例 6.3　$X^2=15\times0.4643^2=3.234<X^2_{\underset{0.05}{(1)}}=3.841$

所求得的 r_{ϕ} 值未达到 0.05 的显著性水平。

4. 积差相关

对于论文式测验题目，因得分具有连续性，在被试团体较大时，可以认为项目分数服从正态分布。可将项目得分与效标分数求积差相关系数以得到项目的区分度。

以上介绍的四种相关法，在实际项目分析中，究竟采用哪一种，依照变量的性质而定。实际上，虽然所得的数值各不相同（$r_{pb}=0.4775$，$r_b=0.599$，$r_{\varphi}=0.4643$），但经显著性检验均未达到 0.05 的显著水平。因此，分析所得的结果是一致的。

三、区分度与难度的关系

在讨论难度指标时，曾提到过测验项目的难度对测验项目的鉴别力有一定的影响，即是说，难度与区分度有着密切的联系。以鉴别度指数 D 为例。例如，某项目的通过率为 1.00 或 0，则说明高分组与低分组全部通过或者没有人通过。此时，两组的通过率没有差异。因此，D = 0。假如题目的通过率为 0.50，则有可能是高分组的所有被试都通过了，而低分组却无人通过，这样 D 的最大值可能达到 1.00。假如项目通过率为

0.70，有可能高分组通过率为1.00，低分组的通过率为0.40，就可使得区分度的值为D=0.60。根据同样方法可求出不同难度的项目可能的最大D值，见表6.3。

表6.3 D的最大值与项目难度的关系

项目通过率（P）	D的最大值
1.00	0.00
0.90	0.20
0.70	0.60
0.60	0.80
0.50	1.00
0.40	0.80
0.30	0.60
0.10	0.20
0.00	0.00

从上表中可以看出，难度越接近0.50，项目潜在的区分度越大，而难度D越接近1.00或0时，项目的潜在区分度越小。这也就是人们在常模参照测验中，要求项目保持中等难度的道理之一。

为了使项目具有较高的区分能力，似乎应该使所有的项目都保持在0.50的难度最为理想，但是在实际编制测验时，我们却不能要求这么做。因为一个测验中的项目大多趋向于与有关的内容或技能具有某种程度的相关。假若所有的题目都完全相关（r=1），并且都是0.50的难度水平，在一个项目上通过的人在其他各项目上也会通过，在一个项目上失败的人，在其他项目上也将失败，那么一半被试将通过每一个项目。另一

半被试将全通不过。在这种情况下，测验将只有两种分数，满分与零分，成V型分布。这样，从整体来说，测验所提供的信息便相对减少。事实上，如果测验的所有项目都是中等难度，只有在项目的内在相关为0时，整个测验分数才产生正态分布。实际测验中，一般各项目之间都具有某种程度的相关，考虑到这一点，我们在利用项目分析选择试题时，应使项目的难度分布广一些，梯度大一些，使整个测验的难度分布呈正态分布，且平均水平保持在0.50左右。这样才能把各种水平的人都区分开来，并且区分得比较细。

四、区分度的相对性

一般来说，难度是相对而言的，它与测验编制者的技术经验、测验内容、被试团体、统计计算方法等有关。同样，项目的区分度也是相对的，通常与以下几方面有关。

（一）不同的计算方法，所得区分值不同

这点从前面所举的例题6.1、6.2、6.3就可以看出，同样是运用相关法、采用不同的计算公式，所得数值不尽相同（$r_{pb}=0.4775$　$r_b=0.599$　$r_\varphi=0.4643$），鉴于此，在分析同一个测验时，各个项目的区分度值要采用同一种指标，否则不便分析比较。

（二）样本容量大小影响相关法区分度值的大小

一般说来，样本容量越小，其统计值越不可靠。所以在计算出r值后，不能仅从数值大小判断试题的优劣。而应运用统

计显著性检验法，检验区分度值是否显著。

(三) 分组标准影响鉴别指数值 (D)

极端组划分的标准不同，求得的区分长值也不同。分组越极端，其 D 值越大。通常取 27%作为极端分组划分的标准。

(四) 被试样本的同质性程度影响区分度值的大小

被试团体越具有同质性，即个体之间水平越接近，其测题的区分度值就越小。反之，若是施测于具有较大异质性的被试团体，即使是对另外一同质团体来说区分度很小的项目，也可能具有很高的区分度。另一方面，区分度也是相对于不同水平的被试团体的。例如，用于测量初二年级教学水平的试题，对于小学生或大学生来说，均不可能有较高的区分度。所以，项目的区分度大小是针对特定团体而言的。

根据以上 4 点讨论，我们在评价项目的有效性时，应考虑到测验的目的、功能以及被试团体的总体水平，不能将区分度值作为筛选试题的绝对标准。表 6.1 所提供的标准只不过是在编制测验时的一个参考标准而已。

第三节　猜测问题与猜测率

一、客观测验题中的猜测问题与猜测率

在客观题中有一个重要问题是：测验分数确实反映了被试

的真实状况，还是因为猜测而获得成功。因为在客观题中，猜测会提高他们的分数，在是非题，配对题及选项较少的选择题，这种影响格外明显。当被试确实不知道正确答案，而每个选项又具有同样的吸引力，被试凭猜测选择正确答案的机会是$\frac{1}{K}$（K是每题中选项的数目）。这样对是非题（K＝2）而言，猜测就能获得50%的成功机会；而四重选择题，其猜测正确的概率就为25%。显然，大量的猜测就会对是非题和选择题的分数产生很大的影响，从而对测量带来误差，即猜测误差。猜测误差来源有：①猜相对于不猜引起的误差，如果有100道四重选择题的测验中（设每题1分），甲、乙两学生都能正确回答60题，两个人的实际水平相等。若甲生不仅回答确有把握的60题，而且对不会的40题全凭猜测做出选择；而乙生只回答已掌握的60题，对不会的不作猜测。在四重选择题中，对答案猜测成功的概率为$\frac{1}{4}$，那么甲生平均能猜对10题，可获70分，而乙生只得60分。在这个假想的例子中，猜与不猜所导致的差异平均将达到10分左右。②是否猜得对引起的误差。即猜测过程中因随机得分情况不同所引起的误差。按照概率原理，是非题猜对的概率是$\frac{1}{2}$，四重选择题是$\frac{1}{4}$，五重选择题是$\frac{1}{5}$，但这是对被试团体平均而言的，即N个被试参加测验，100个四重选择题仅凭猜测能猜对25题。具体到某一个人，他实际猜对几题并不一定与概率值相等，上面所举的例子中，甲生可能猜对10题而得70分，也可能猜对8题而得68分或猜对12题得72分。这是由猜测本身引起的误差。通过以上两点讨论，有人认为，由于对某些测验项目，猜测会引起项目难度的变化，允许猜测将使通过率或得分高于被试的实际水

平。为此，有必要对猜测进行校正。

二、项目难度受猜测影响的校正

（一）猜测校正的性质与公式

在选择题测验中，猜测的成功概率受项目备选答案数目（K）的影响（$P=\frac{1}{K}$），备选答案数目越少，机遇的作用越大，被试的得分将越高于他们的真实水平，根据难度的计算公式（6.1）、（6.2）求出的难度的就越不能反映出项目的真实难度。为平衡机遇对难度的影响，可采用下式来对难度进行校正：

$$CP=\frac{KP-1}{K-1} \tag{6.11}$$

$$\text{或}\quad CP=P-\frac{q}{K-1} \tag{6.11'}$$

式中 CP 为校正后的通过率，P 为实际通过率，K 为备选答案数目；$q=1-p$。

如果要比较两个选项数目不同的测题难度，必须应用公式（6.11）分别将两个测题的难度进行校正，然后才能进行比较分析。

例 6.4 有 A、B 两个测题，项目 A 为四重选择题，通过率为 0.58；项目 B 为五重选择题，通过率为 0.56；试比较两题的难度。

解：采用公式（6.11）对难度进行校正，消除猜测因素的影响。

对项目 A：$CP=\frac{KP-1}{K-1}=\frac{4\times0.58-1}{4-1}=0.44$

对项目 B：$CP=\frac{KP-1}{K-1}=\frac{5\times0.56-1}{5-1}=0.45$

如果根据未经校正的难度相比，A 项目比 B 项目容易（0.58 > 0.56），根据校正后的难度相比，B 项目比 A 项目还稍容易些（0.45 > 0.44），其结论正好与校正前相反。可见在这种情况下，必须经过校正后，才能进行比较。因为选项数目不同的选择题，受猜测机遇的影响大小不同。

公式（6.11）是对全体被试而言的，即根据被试团体在某项目上的通过率而计算校正难度。若对某个被试来说，参加由多个项目所组成的测验，同样有必要对他们的得分进行校正，以求出能反映出他真实水平的校正分数，校正公式只需将公式（6.11）稍作变换即到下式：

$$S=R-\frac{W}{K-1} \tag{6.12}$$

式中 S 为校正后的得分，R 为被试答对的项目数，W 为被试答错的项目数，K 为项目的选项数目。

例如，某被试参加由 100 道四重选择题组成的测验，测验结果是答对 82 道题，答错 18 题，该被试的实得分数为（每题 1 分）$S=82-\frac{18}{4-1}=76$（分）。其理由在于四择一选择题中，每题猜对的概率为$\frac{1}{4}$，猜错的概率为$\frac{3}{4}$，该被试答错 18 题，说明他猜测了 24 道题（$24\times\frac{3}{4}=18$），其中猜对 6 题（$24\times\frac{1}{4}=6$）。因此，实际确能掌握的只有 76 道题。

（二）猜测校正的优缺点

公式（6.11）、（6.12）的基本假设是：被试不知道正确答案时，完全凭猜测作答，猜测的成功与否完全由随机因素所致，即选择哪一个备选项是随机决定的。在实际测验中，这种

假设很少成立。因此，对上述校正公式存在很大的分歧。

赞成使用猜测校正公式的人认为：

(1) 可避免降低测验的信度。因为如果不使用校正公式，被试必然会盲目猜测而影响测验的信度；使用答错题倒扣一定的分数，则被试不敢盲目猜测。

(2) 校正后的得分可以反映被试的真正水平和能力。对每个项目来说，校正后可以反映项目的真实难度，便于在备选答案数目之间进行统计比较分析。

(3) 在教育测验中，可以培养被试诚实的美德。如果鼓励尽量答题，并允许猜测，且不扣分，则会使学生心存侥幸，有害于健全人格的培养。反之，如果采取校正猜测，则可养成学生"知之为知之，不知为不知"的良好品德。

(4) 比较公平。即使事前鼓励学生答完全部试题，但事实上总有人无法答完全部试题，所以使用猜测校正的方式比较公平。

反对使用猜测校正公式的人认为：

(1) 公式的基本假设不成立。因为被试答错试题，并非都是存心投机取巧。事实上，有些学生答错，可能是观念模糊、记忆错误或粗心大意所致。大多数情况下，均是先舍弃部分诱答，再就剩下的几个选项来猜测，而非盲目猜测。

(2) 只要被试能答完全部试题，则猜测校正无实质作用。根据统计学方法，将分数转化为相对分数后，校正前后的分数完全相同，说明两种分数对于决定分数的高低具有相同的作用。虽然校正前后分数不同，但两者的相关系数为 1.00，所以采用校正，只是采用线性变换，降低被试的得分，增加记分的复杂性，不仅浪费时间，且易发生错误。

(3) 不采用猜测校正对信度并无重大影响。根据台湾学者黄国彦研究 (1977)：鼓励被试猜测，其影响只有 4%左右，此项缺点可通过增加试题的数目来提高测验信度。

(4) 有时会出现无法解释现象。一个学生如果答对的题数等于或少于答错的题数。校正后便会得到零分或负分。例如，在一个有100道四重选择题的测验中，某被试答对22题、答错78题（每题1分），此人校正后的分数为 $S = 22 - \frac{78}{4-1} = -4$（分），这是难以解释的。因为按普通常识来说，即使一点都不了解测验所测的知识内容，也不过得零分，而不至于得负分的。

(5) 实际生活中，经常缺乏充分的证据与资料，必需凭借部分知识来判断，且进行合理猜测是值得培养的习惯。测验时，若不准被试猜测，则与现实生活情况不符。事实上，许多科学上的发现是在把握不很大的情况下，先提出猜想，而后慢慢证实的。人的某些猜测依靠的是直觉思维，这是对事物整体的认识，虽然没有经过严密的逻辑推理，但并非完全瞎猜。

综上所述可知，对于是否需要采用猜测校正，并无定论。但是在答题时间充裕，备选答案数目（K）在四个或以上的选择题，则没有必要进行校正记分。

第四节 多重选择题的项目分析

多重选择题因能比较有效地控制随机猜测导致的测量误差，能测量较复杂认知目标，能为改进教学提供更多的反馈信息，且具有易于评分、能用计算机阅卷等优点，在教育与心理测验中，应用极其广泛。对于多重选择题，当然可以采用本章第一、二节所介绍的内容，进行难度与区分度分析。为了进一

步提高测验质量，充分发挥选择题的功能，除了进行难度区分度分析之外，还应对被试在项目作答反应上进行分析。对多重选择题作项目分析，可以解决以下问题。

(1) 项目是否具有所预期的功能？对于常模参照测验，测题是否有足够的区分度？对目标参照测量来说，测题是否能充分地测量到教学的结果？

(2) 项目的难度是否得当？

(3) 项目是否有缺陷？

(4) 诱答选项是否都有效？

对于(1)、(2)两点，可采用本章第一、二两节所介绍的方法进行分析。本节主要就(3)、(4)两点进行讨论，即通过被试对选择项反应模式的分析来改进并提高测验项目的质量。

具体分析的步骤：

(1) 按被试测验的总分，从高到低依次排列试卷。

(2) 从最高分依次向下取全部试卷的27%作为高分组。

(3) 从最低分依次向上取全部试卷的27%作为低分组。

(4) 分别登记高分组与低分组选中各选择项的人数（亦可将人数换为人数比例），然后登记。

(5) 根据登记结果进行选择项的质量分析。

对选择项的反应模式注意从以下几方面进行分析：

(1) 如果正确的备选答案被所有的受测者所选择，说明该项目太容易或者可能是项目中提供某种暗示，使正确答案过于明显。

(2) 如果某个错误答案没有任何被试选择，则说明该选项不具有迷惑性，错得过于明显，除增加阅读时间外，不起任何作用。一般说来，除非有2%以上的人的选择，否则该备选答案应该修改或删掉。

(3) 如果所有被试都选择了同一个错误答案，可能是编制测验时把正确答案搞错了，也可能是在教学中发生了错误。

(4) 如果高分组被试的选择集中在两个答案上，二者选择率相近，说明该题可能本来就有两种正确答案，或者在某种意义上另一个选择项也有一定的道理。

(5) 如果高分组对正确答案的选择率与低分组相等或低于后者，说明该题所考察的东西与被试水平无关，即不具有鉴别力，此题应删除或作大的修改。

(6) 如果一个题目被试未作答的人数较多（速度性测验除外），或选择各个备选答案的人数相等，说明该项目可能过难或题意不清，被试无法作答或凭猜测作答。

在实际进行分析时，可以将多重选择题的选答情况登记在一张选择分析表中，以便于进行分析评价。兹举例如下。

例 6.5　下表为一个由 370 人参加的测验中的 4 道题的项目统计结果，据此表对此四题作分析评价。

题号	组别	选答人数					正确答案	难度	区分度	
		A	B	C	D	未答		P	r_{pb}	D
1	高分组	5	92	1	2	0	B	0.71	0.52	0.42
	低分组	22	50	12	16	0				
2	高分组	58	10	15	16	1	A	0.42	0.33	0.32
	低分组	26	21	15	36	2				
3	高分组	17	25	28	28	12	D	0.31	-0.04	-0.06
	低分组	25	11	19	34	11				
4	高分组	1	44	14	36	5	C	0.12	0.08	0.04
	低分组	1	56	10	28	5				

该表中的高分组、低分组是按测验总分的高低，从 370 人中按 27%的人比例选取的。

(1) 难度。第一题的难度较小，第二题难度适中，第三、

第四两题难度较大。

(2) 区分度。第一、第二两题的区分度符合要求具备良好的测题的首要条件，第三、第四两题的区分度不够，第四题太小，而第三题则是负向的，此两题均为不良试题。

(3) 各题的选项分析。第一题：正误答案配比较好。第二题：除C答案缺乏鉴别能力外，其余都不错。但值得注意的是为什么在C答案上，高低分组的选答人数相同。第三题：未答的人数比例过大，且答案B、C、D均属于负向的，高分组选C的人数较多，等于选正确答案D的人数，这些均要研究。第四题：A项选答的人数较少，是否因为该选项错得太明显而缺乏似真性，另外D选项也有负向性，须找出原因加以适当修改。

练习与思考

1. 测验项目分析的作用是什么?

2. 某测验对11名被试施测，结果数据如下表，试计算各题的难度、区分度。

题号	满分值 \ 考生	A	B	C	D	E	F	G	H	I	J	K
1	3	3	3	0	3	3	0	3	0	0	3	3
2	5	5	0	5	5	0	0	5	5	5	0	0
3	10	8	8	5	9	10	3	7	10	10	5	7
4	12	10	12	7	8	5	5	9	8	7	6	7
5	20	15	10	12	17	15	10	15	17	18	15	10
6	50	45	30	20	42	35	25	38	38	44	40	23
合计	100	36	63	49	84	68	43	77	78	84	69	50

试计算各题的难度、区分度。

第七章　测验常模

本章提要：

- 各种常用导出分数及其之间的关系
- 各种测验分数合成的方法
- 常模编制及常用常模

从测验中直接获得的分数，称为原始分数。它是通过将被试的反应与标准答案相比较而获得的。但是原始分数本身并不具有多大意义，在实际应用中，需要配以可供比较的标准，将原始分数转换以得到有意义的、可供解释的分数。这种比较标准就是由原始分数的分布转换过来的具有参照点和单位的测验量表。本章首先介绍几种常用的分数转换方法与导出分数，然后讨论常模量表的制定过程和方法，并介绍几种主要的常模参照分数。

第一节 分数转换

一、原始分数与导出分数

被试在接受测验后，根据测验的记分标准，对照被试的反应所计算出的测验分数称作原始分数。原始分数反映了被试答对题目的个数或作答正确的程度。但是原始分数一般不能直接反映出被试之间的差异状况，不能刻划出被试相互比较后所处的地位，也不能说明被试在其他等值测验上应获得什么样的分值。为了使原始分数本身具有意义，使不同测验的分数可以相互比较，就必须将原始分数转换为导出分数。

导出分数就是在原始分数转换的基础上，按照一定的规则，经过统计处理后获得的具有一定参考点和单位，且可以相互比较的分数。这种按某种规则将原始分数转化为导出分数的过程称作为分数的转换。常用的导出分数有百分等级、标准分

数、T分数等。

二、百分等级分数

（一）百分等级分数的概念

百分等级是应用最广的导出分数。一个原始分数的百分等级是指在一个群体的测验分数中，得分低于这个分数的人数的百分比。也就是说，如果将某一被试群体分为一百个等级，则每位被试所占的等级数就是百分等级。例如，某一被试在一项测验中得82分，经过换算，百分等级分数为75，就表示参加该项测验的人得分低于82分的占全体被试的75%，并说明超过他的成绩82分的人仅有25%，我们通常用P_R来表示百分等级。显然，百分等级取值越大，说明成绩越优秀。

（二）百分等级分数的计算

1. 未分组分数资料

对于未分组分数资料，求一个原始分数的百分等级，可先将被试团体的全体原始分数从大到小排序，然后采用下列公式计算：

$$P_R = 100 - \frac{100R - 50}{N} \qquad (7.1)$$

式中，P_R为百分等级，R为排名顺序的序号，N为被试总人数。

例如：某被试在一次由50人参加的成绩测验中得80分，排名第9，则该生成绩（80分）的百分等级为：

$$P_R = 100 - \frac{100R - 50}{N} = 100 - \frac{100 \times 9 - 50}{50} = 83$$

其百分等级为 83，即是说比 80 分低的原始分数占全体得分的 83%，比其高的只占 17%。

2. 分组分数资料

如果被试团体较大，往往已对分数作过初步整理，分数资料通常以次数分布表的形式呈现，此时，可采用下列公式求得百分等级。

$$P_R = \frac{100}{N}\left[\frac{(X-L)\ f}{i} + F_b\right] \qquad (7.2)$$

式中 X 为被试原始分数，L 为 X 所在组下限，f 为 X 所在组的次数，F_b 为 X 所在组以下各组次数之和，i 为组距，N 与 P_R 同（7.1）式解释。

例 7.1　一次由 250 人参加的数学测验，分数经整理，分布情况见表 7.1，某被试得分为 78 分，试求其百分等级。

表 7.1　250 名学生数学测验原始分数次数分布

分数（X）	次数（f）	累积次数（f）
95 ~ 100	3	250
90 ~ 95	11	247
85 ~ 90	18	236
80 ~ 85	27	218
75 ~ 80	49	191
70 ~ 75	65	142
65 ~ 70	38	77
60 ~ 65	25	39
55 ~ 60	13	14
50 ~ 55	1	1
$\sum$	250	——

解：先求向上累积次数（列于表中第 3 列）

由上表可知：X = 78，N = 250，i = 5，L = 75，F_b = 142，

代入公式（7.2）得：$P_R = \frac{100}{N}\left[\frac{X-L}{i}\cdot f + F_b\right] = \frac{100}{250}\left[\frac{78-75}{5}\times 49 + 142\right] = 68.56$

百分等级往往按四舍五入原则取为整数，故该生的百分等级为69。

（三）对百分等级分数的评价

百分等级是一种相对位置量数，具有可比性，且具有易于计算、解释方便等优点，对一般教师、学生和家长来说，均能了解百分等级的意义，所以它较适用于不同的对象和性质不同的测验。另外，百分等级不受原始分数分布状态的影响，即使分数分配不是正态的，也不会改变百分等级常模的解释能力。

但百分等级是一种顺序量数，它在统计分析中不具有可加性。在实际应用中，有以下两个缺点：

（1）单位不等，尤其在分配的两个极端。如果原始分数的分配是正态或近似正态分布，则靠近中央（平均数或中位数附近）的原始分数转换成百分等级时，分数之间的差异便夸大了，虽然原始分数比较靠近，但转换成百分等级后，却显示出很大的差异性；对接近两极端的原始分数，百分等级反应迟钝，即使原始分数发生较大的变化，也不能引起百分等级的相应变化，使得其差异被缩小了。例如表7.1的资料中，原始分数60、65和70、75两对数据，虽然原始分数之差均为5分，但转换为百分等级后，其差异就会有很大的区别，前者只差10.4个百分等级（15.6%～5.2%），而后者则相差26个百分等级（56.8%～30.8%），两者有明显区别。

（2）百分等级只具有顺序性，而无法用它来说明不同被试之间分数差异的数量。例如，某被试甲在一个成就测验中的百分等级为10，被试乙为20，被试丙为30，我们只能说丙优于

乙，乙优于甲，而不能推断他们之间差异的程度相等。因此，百分等级不适合计算平均数、相关系数及其他统计量数。

另外，在运用百分等级时应注意到，百分等级是相对于特定的被试团体而言的，所以，解释时不能离开特定的参照团体。被试得分不变，但参照团体改变了，百分等级值就可能发生变化。例如，某被试成绩 80 分，以他所在班为参照团体，可能处在第 75 百分等级上，但若以全年级为参照团体，其百分等级就不一定是第 75 百分等级了。如果他所在班本次测验在全年级中是优秀的班，则他的百分等级值可能会高于 75，反之，若是差的班则就会低于 75。所以在报告百分等级时，一定要说明是相对于什么的参照团体来说的。

三、标准分数

（一）标准分数的意义

标准分数是一种具有相等单位的量数，又称作 Z 分数，以 Z 表示。它是将原始分数与团体的平均数之差除以标准差所得的商数，是以标准差为单位度量原始分数离开其平均数的分数之上多少个标准差，或是在平均数之下多少个标准差。它是一个抽象值，不受原始测量单位的影响，并可接受进一步的统计处理。

（二）标准分数的计算

标准分数的计算公式为：

$$Z = \frac{X - \bar{X}}{S} \qquad (7.3)$$

Z：标准分数；X：原始分数；$\bar{X}$：团体所有被试的原始分数的平均数；S：原始分数的标准差

（三）对Z分数的评估

Z分数是以一批分数的平均数为参照点，以标准差为单位的等距量表。Z分数不仅具有可比性，而且还具有可加性，它由符号与绝对值两部分构成。正负符号表示原始分数在平均数之上或之下，绝对值表示原始分数与平均数的距离。除此之外，Z分数还具有以下两个重要性质：

（1）运用（7.2）式所求得的Z分数，实际上只是对原始分数X所作的一个线性变换，所以Z分数与原始分数X的分布形态相同。若原始分数不服从正态分布，转换成Z分数后，其分布仍然非正态。

（2）任何一组原始分数经转换为Z分数以后均有$\bar{Z}=0$、$S_Z=1$，所以可以利用Z分数对不同测验分数进行比较。如果原始分数属正态分布或近似正态，则Z分数的范围大致在-3.00到$+3.00$之间（约占全体的99.73%）。

但Z分数由于计算中经常出现负数和小数，且单位过大（一个标准差单位），所以，使用起来不够方便。

（四）正态化的标准分数

将原始分数转换成导出分数的原因之一，是为了对不同测验中的分数进行比较。但是Z分数与原始分数的分布形态相同，所以只能在两个原始分数分布形态相同或近时才能运用Z分数进行比较，否则，仍然无法作直接比较分析。比如说，若两个分布的偏斜方向不同，或一个正态、一个为偏态，则相同的Z分数可能代表不同的百分等级，对于这两个测验分数，仍然无法准确比较。为了使来源于不同分布的分数进行比较，

可使用非线性变换，将非正态分布的分数强制性地扭转成正态分布。具体做法为：首先将每个原始分数转换为百分等级，然后使用正态分布表，将对应的百分等级直接看成是正态分布曲线下的面积值，找出所对应的 Z 值（偏差值），这种方式所得到的分数叫作正态化的标准分数。图（7.1）即为负偏态分布转换为正态分布的示意图。

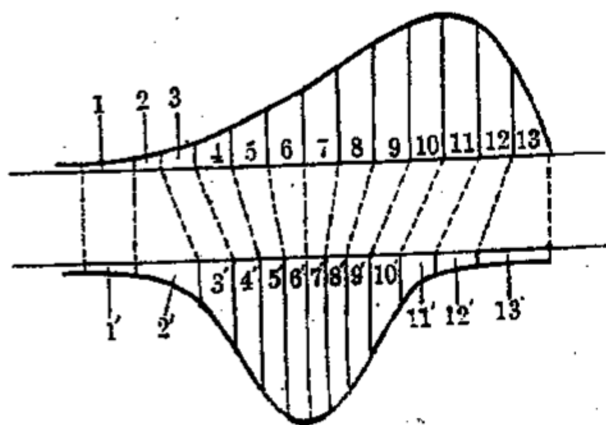


图 7.1　负偏态分布正态化图

四、标准分数的变式

（一）T 分数

1. T 分数的意义

由于 Z 分数常常带有小数和出现负值，使用起来常觉不

便，也容易出错，并且与日常生活中的评分形式不一致，不直观。因此，产生了多种将 Z 分数作线性变换，使负号与小数消失，全部变为正数的转换方法。最早由美国测量学家麦柯尔建议（1939）将 Z 分数扩大 10 倍（以消除小数）再加上 50（消失负号）。为纪念推孟与桑代克，这种转换后的分数命名为 T 分数。所以 T 分数实际上是由标准分数直接转换而来。后来，人们在麦柯尔思想的基础上，又衍生出多种导出分数。

2. T 分数的计算

最初，麦柯尔所采用的 T 分数为：

$$T = 10 \cdot Z + 50 \qquad (7.4)$$

式中，T 为 T 分数，Z 为标准分数。

麦柯尔的 T 分数是对单科标准分数的变换，T 在〔0，100〕之间，T 分数的平均数为 50，标准差为 10，T 分数避免了小数与负号。但如果原始分数服从正态分布，转换后的 T 分数，就有一半的人在 50 分以下，若不加区别地当成百分制分数使用，并简单地以通常采用的 60 分为及格线，势必就有 83% 以上的被试不及格，则与日常教育测验中分数的解释就相悖了。

（二）其他形式

按建立 T 分数的思想，在 Z 分数的基础上，进行线性变换，导出了多种适合不同需要的标准分数形式。其通式为：

$$Z' = A \cdot Z + B \qquad (7.5)$$

式中 Z′为由 Z 导出的导出分数，A、B 为常数。

常见的变化形式有：

（1）美国大学入学考试委员会使用的标准分数，即 CEEB 分数，公式为：

$$\text{CEEB 分数} = 100 \cdot Z + 500 \qquad (7.6)$$

平均分数为 500，标准差为 100。

(2) 韦氏智力测验采用的离差智商，转换公式为：

$$IQ = 15 \cdot Z + 100 \tag{7.7}$$

IQ 平均为 100，标准差为 15。

(3) 我国一种出国人员英语水平考试即 EPT 所使用的分数转换公式为：

$$\text{EPT 分数} = 20 \cdot Z + 90 \tag{7.8}$$

平均分数为 90，标准差为 20。

(三) 标准分数变式的评价

以上介绍了几种常见的标准分数变化形式，它们都是以 Z 分数为基础进行线性变换而来。它们具有以下几点优点：

(1) 具有等单位特点，便于工作进一步的统计分析。

(2) 正态分布下，可以利用正态分布表将各种导出分数与百分等级分数作换算；

(3) 正态分布下，运用某种变式分数可以将几个测验上的分数作直接的比较。即使是非正态分布，也可运用由正态化的 Z 分数转换而得的变式分数进行直接比较分析。

关于变式分数的缺陷，主要归纳为以下几点：

(1) 分数过于抽象，不易理解，正如在介绍麦柯尔的 T 分数时所提到的那样不为一般人所熟悉。

(2) 在非正态分布下，分布形态不同的变式分数，仍然不可以作相互比较，也不能相加求和。

五、标准九分数

标准九分数是将原始分数分成几个部分的标准分数系统。

若原始分数服从正态分布，它是以0.5个标准差为单位，将正态曲线下的横轴分为九段，最高一端为9分，最低一端为1分，中间一段为5分，除两端（1分，9分）外，每段均有半个标准差宽。在正态分布下，每个标准九分所占的位置与包含的百分比如表7.2所示。

表7.2　标准九分与正态分布的对应关系

标准九分	本段变积	累加变积	本段中值与平均数的距离
9	4%	100%	大于2.0σ
8	7%	96	1.5σ
7	12%	89	1.0σ
6	17%	77	0.5σ
5	20%	60	0.0
4	17%	40	0.5σ
3	12%	23	1.0σ
2	7%	11	1.5σ
1	4%	4	大于2.0σ

如果原始分数分布不是正态的，只要将原始分数转换成为百分等级就可以很容易的从表7.2中求得被试的标准九分数。譬如，某被试的原始分数在团体中处于第75百分等级，则由表7.2可推知该被试的标准九分为6分。

六、几种导出分数间的相互关系

在教育与心理测量中，由于被试群体较大，所测特质的得分分布形态一般都能保持正态或近似正态。在正态分布下，各种导出分数之间的关系如图 7.2 所示。

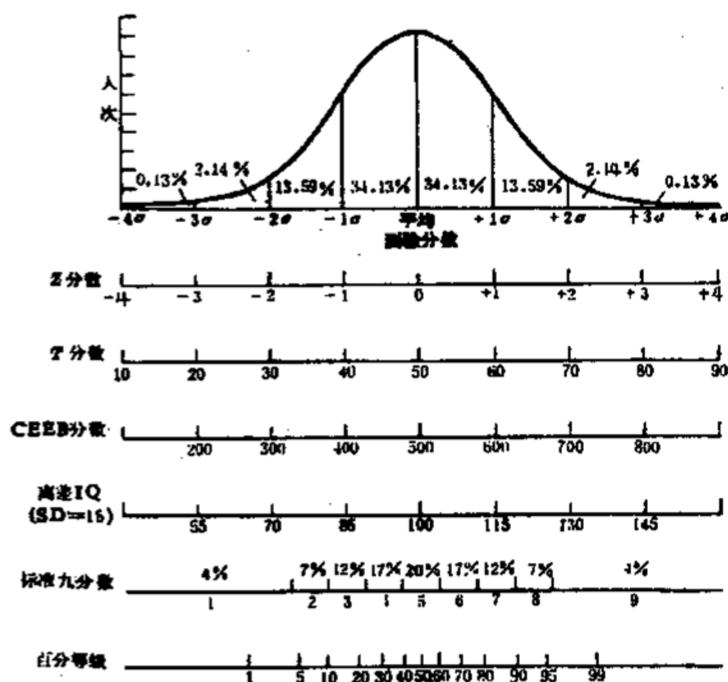


图 7.2 常用导出数的对应关系

第二节 分数合成

一、分数合成的意义

（一）分数合成的种类

前面所介绍的分数转换，通常都是对一个测验的分数而言的。实践中只处理单一测验分数的情况很少，常常需要将几个分数或几个预测源组合起来，以获得一个合成分数或作总的预测。例如，高等学校录取新生，不仅需根据多科学业成绩的得分情况，还要结合思想表现与体检结果等多方面测验结果进行整合，择优录取，择优的标准事实上就是将多方面得分合成后所得的结果。我们常遇到的组合有3种类型：由基本测验项目组成一个分测验或一个测验；由几个分测验上的得分组成合成分数；由几个测验的得分组合，获得合成分数或合成预测。

（1）项目的组合。每个测验是由许多独立的项目所组成。这些项目可以结合成小组，各小组的项目可以独立组合成量表或分测验，也有直接将所有项目得分合成一个测验总分的。在这种情况下，总分均为个别项目得分的合成分数。虽然大部分的分数是对所有项目等量加权而得到的合成体，但个别题目也可以作不等量加权。不论是否采用加权方法，除非测验使用者对个别项目具有特殊兴趣，否则通常均要把各个项目分数合成以得到测验总分。

（2）分测验或量表的组合。有些测验是由几个分测验或分

量表所组成，每个分量表均有个分数，这些分数可以组合到一起得到一个合成分数（当然有时也可以不这样做）。例如韦克斯勒成人智力量表由言语量表与操作量表两部分构成，而言语部分包括6个分测验，其合成分数叫作言语智商，操作部分包括5个分测验，合成分数为操作智商，还可以将11个分测验总合成而得到总智商。

（3）测验或预测源的组合。在作实际决定时，常常将几个测验或预测源同时使用。如前面所提的大学录取新生，即是将各科测验分数与其他成绩合成后作为录取依据的。又如美国雇佣服务中心，对申请者实施几个测验，测量9个因素，用来预测在各种职业上的成功。以上两例，均是测验使用者为了作出决定而将测验分数进行某种组合的。

（二）分数合成中的问题

每当将测验分数组合时，必须考虑以下3个问题：

（1）采用什么方法来合成分数？考虑该问题主要取决于组成测验分数的目的与要作何种决定。如果分数合成后根本不能为实现测验目的服务，就没有合成的价值。另外，测验所测特质间能否彼此替代，测验所获资料的性质以及整个工作的效率与效益如何，都对测验分数合成方法的确定有一定的影响，我们需考虑到这些因素，以便选用科学、有效、经济的方法。

（2）什么形式是最适当的分数组合？这个问题基本上是效度问题。一般而言，我们只对能产生最高效标效度的测验组合感兴趣，所以，可用效标效度来评价合成分数。但是，如果在效标效度不是我们最关心的问题的情况下，也可用其他标准来评价。

（3）需要多少及何种测验分数作最适当的组合分数？组合分数时，使用的测验分数的种数即测验的个数并不是越多越

好。假如使用3个测验组合成的分数与使用6个测验组合成的分数的效果大体相当，我们自然是只使用3个测验。通常当将测验组合，用来预测一个效标时，以最好的一个预测源开始，然后再添加预测源，直到组合分数的效度不再增加为止。若一个测验加入测验组合体而没有使效度增加，则表示该测验并没有提供任何新的信息，就不必增加。

二、分数合成的方法

在讨论各种分数合成方法时，将不区分是组合各个项目分数、分测验分数或测验分数，因为不论以何种单位分析，其原理是一致的。

由于测量目的和所用资料不同，组合方法既可以是统计的，也可以是推理或直觉的。

（一）临床诊断——直觉合成

在实际工作中，最常用的组合测验分数的方法是根据经验对测验分数作直觉的组合，这就好比临床医生，把各种化验、检验所获得的资料与实际观察所得的结果结合起来，根据经验作出诊断一样。与此相似，一个教师或家长在帮助学生填报高考志愿、选择大学和学业时，根据该生的平时成绩、高考各科估分、兴趣爱好、专长性格及招生情况等各种因素，全面分析并作出判断。像这种根据直觉的经验，主观地将各种因素加权，而获得结论或预测的方法叫作临床诊断。

临床诊断法的优点是：①具有高度的综合性。它允许我们从整体上来考虑问题，充分考虑各测验所测特质间交互影响，

各测验上所得分数的对比关系与组合类型的结构特点，测验分数与实际反应表现其中的生动关系等。②具有灵活的针对性，能就特定的个人作具体的结论。而一般的统计方法具有常模性，常模性的统计模式难于适应每个个体所具有的独特性，更难于适应非典型的新颖形式。

临床诊断法的缺点是：①主观加权易受决策者的偏见影响，不够客观。②缺乏精确的数量分析，没有精确的数量指标。

（二）加权求和合成

如果各个测验所测特质间有相互代偿作用，这些测验上的分数又是连续性资料，并能大体同时获得（如学生各种考试成绩），那么可以采用加权求和的立法对分数进行合成。

最简单的加权方法为单位加权，就是将各个测验分数直接相加而获得合成分数。

即：
$$X_c = X_1 + X_2 + \cdots + Xn \tag{7.9}$$

式中 X_c 为合成分数，$X_1 \cdots Xn$ 为各分测验分数，以往高考总分就是采用这种方法将各科分数作单位加权而获得的。

虽然（7.9）式看起来好像对所有变量作了等量加权，事实上，这方法是根据每个变数与它的标准差成比例的加权，即将变异量最大的测验作最重的加权。假如想将变量作等量加权，可以将所有测验分数转换为标准分数，然后采用下式加权组合。

$$Z_c = Z_1 + Z_2 + \cdots + Z_n \tag{7.10}$$

式中，Z_c 为合成的标准分数，Z_1，$Z_2 \cdots Z_n$ 为各分测验的标准分数。（7.10）式适合于各测验对预测效标具有同等重要性的场合。但在通常情况下，各个变数对预测效标的作用是不同的。因此，需要根据各个变数与效标之间的经验关系作差异

加权。其通式为：

$$Z_c = W_1Z_1 + W_2Z_2 + \cdots + W_nZ_n \qquad (7.11)$$

式中 Z_c，Z_1，Z_2，…，Z_n 同（7.10）式，W_1，$W_2 \cdots W_n$ 是加权系数。

加权系数的确定比较复杂，通常采用的方法有：①抽象推理，从某些理论要求出发加以推定。②使用统计学方法，常用主成分分析的第一主成分作权数，读者可参考有关统计著作。

（三）多重回归

采用加权合成所得到的分数，是各个分测验分数的综合值，但在很多情况下，需要利用测验结果对预测效标作出估计。例如，根据高考各科成绩预测在大学一年级末的学业成绩等。此时，需对测验结果和效标测量作多重回归分析，求出效标估计与预测变量之间的数量关系式。

多重回归就是研究一种事物或现象与其他多种事物或现象在数量上相互联系和相互制约的统计方法，基本方程式为：

$$\hat{Y} = a + b_1x_1 + b_2x_2 + \cdots + b_nx_n \qquad (7.12)$$

式中 $\hat{Y}$ 为预测效标分数；x_1，x_2，…，x_n 为各个预测源分数；$b_1 \cdots b_n$ 为每个预测源的加权数；a 为一常数，用来校正预测源与效标平均数的差异。

多重回归分析的输入资料为预测源与效标的平均数与标准差，以及所有变量间相关的相关距阵。分析过程包括解一系列的联立方程式。通过对预测源作适当的加权而使这些加权的测验分数的合成能以最小的误差来预测效标分数，这些都必须通过电子计算机进行。输出的结果主要有两项：①回归方程式以指出各个预测源的加权量。②复相关系数 R_1 表示预测源(当作一个合成体)与效标测量间的相关。R^2 为决定系数，表示效标中的变异数可由预测源来解释的比例。多重回归的计算原理读

者可参考有关统计学著作。

从理论上说，可以用任何数目的变量来作为预测源。但在实际分析时，则首先用最佳的预测源，即选出与效标相关最高的变量，然后加入另一预测源组合起来以使 R 的数值增至最大，下一个要加入的预测源应该是与前两个预测源组合起来能使 R 值增加最多的，依次类推，当加入额外的预测源不再显著地使相关系数 R 值增加时，则终止分析。在实际应用中，一般二至四个预测源就足以达到最高的预测正确性。但是在具体应用时，应注意到多重回归方法所采用的是统计线性模型，所以只有当预测源与效标间存在线性关系时才是适合的。同时还要求预测源分数跟效标分数能够同时取得，并且都是连续性资料，若这些条件不能满足，就不宜于采用多重回归分析而应用其他方法。

（四）多重划分

用多元回归分析组合分数，适合于所测特质具有某种程度的互偿性。例如高考中，某人某门功课较差，但可以通过其他几门获得高分而弥补缺失，使之可以录取。但实际生活中，有些所测特质之间是不能互相补偿的，例如招收飞行员的筛选，其中任何一项检测不合格者都不能录取。多重划分就是在各个特质上都确定一个标准，从而把成绩划分为合格与不合格两类。在一个测验上合格了，不能保证总的要求一定能合格。只有每个测验都合格时，总要求才算合格。如果有个人几乎在前面所有的测验上得出奇的高分，但在接下来的一次测验中得分低于规定的分数线时，他同样要被删掉。所以在整个测验实施时，是把所有组成这一测验的分测验按一定顺序排列起来逐一实施。只有通过了前一次测验，才能继续实施后一个测验。当有一个测验的成绩被断定为不合格时，测验即停止，被试被视

为不合格而予以淘汰。所以被试要想得到完全合格的结果，就必须使各个测验的分数均达到规定的分数。由于成功的被试必须越过一连串测验的栅栏，所以这种方法也叫作“连续栅栏”。

采用多重划分的方法组合分数时，应该将最有效的预测源或测验放在前面，紧接着为第二个有效的测验，如此类推。这样就能保证整个逐步淘汰过程具有最优良的选择效率。

采用多重划分方法，只决定接受或拒绝，每个被试只可放在其中一类别：达到最低标准与没有达到最低标准。因此，在通过连续栅栏选择的被试之中，相互之间没有优劣之分，他们之间的差异被忽视了，若想区分他们之间水平的差异，必须用其他方法。

以上介绍的几种常用的分数组合方法，在实际运用时，应注意合成方法的适用范围，合理使用。必要时，应将几种方法结合起来，并考虑到有关实际情况，寻求效果最佳且经济实惠的合成方案。

第三节　常模编制

本章开头已指出，常模是根据标准化样本的测验分数经过统计处理而建立起来的具有参照点和单位的测验量表。在这个量表上，被试可根据自己的测验分数找到自己在团体中所处的地位。编制常模需要三步：①确定有关的比较团体。②获得该团体成员的测验分数。③把原始分数转化为量表分数。

一、常模团体与常模

1. 常模团体

常模团体是由具有某种共同特征的人所组成的一个群体，或是该群体的一个样本。

由于个人相对等级随着用作比较的常模团体的不同而有很大的变化，所以任何一个测验可能有许多常模团体。故在制定常模时，首先确定常模团体；在作常模参照分数的解释时，也必须首先就考虑到常模团体的组成。

从测验的编制者来说，确定常模团体的问题，变成确定所编制的测验将来用于什么总体，所选定的常模团体必须能够代表该总体。例如，测验是设计来评价高中毕业生的学业成就的，则常模团体应包括全体高中毕业生，或是能足够代表该总体的一个样本。由于大部分的测验要用于各种不同团体，所以大部分测验都有不止一个常模团体。如瑞文标准推理测验，常模团体就有儿童、成人、城市、农村等多个。对测验的使用者，要从不同角度来选定常模，首先要考虑的问题是现有的常模团体哪一个最适合？因为标准化测验通常提供许多原始分数与各种常模团体的比较转换表，被试的分数必须与最合适的常模比较。

无论是测验编制者还是测验使用者，所关心的主要问题仍然是常模团体的成员。对于成就测验和能力倾向测验，适当的常模团体通常包括目前与潜在的竞争者；比较广泛的能力与性格测验，常模团体通常包括具有同样年龄或教育水平的人。当然，在一些特殊情况下，还有许多方面也可用来定义常模团

体，如性别、年龄、年级或教育水平、职业、社会经济地位、民族等。

2. 确定常模团体的注意事项

（1）群体构成的界限必须明确。在确定常模团体时，必须清楚地说明所要测量的群体的性质与特征。虽然有关常模团体的一般规定取决于测验的目的与使用，且可能有多个常模团体。但对每个常模团体的性质和特征必须有一个简短而明确的描述，若群体过大，群体内部也许有许多小团体，它们在一个测验上的表现也时常有差异，假如这种差异较为显著，就必须对每个小团体分别建立常模。例如，艾森克个性预测（EPQ），就是分性别，以不同年龄组而建立常模的。

（2）常模团体必须是所测群体的一个代表性样本。当所要测量的群体较小时，将所有的被试逐个测量以得到常模。在群体较大时，则不可能如此，只能测量一部分被试作为群体的代表，此时就存在取样是否具有代表性的问题。如果常模团体缺乏代表性，将会使常模资料产生偏差，从而影响到测验结果解释的准确性。为了克服取样偏差，保证具有代表性，一般在抽样时应遵循随机化原则，采用统计学的方法抽取样本。关于具体抽样方法，可参阅有关统计学著作中的抽样推断部分。

（3）取样的过程必须明确且有详尽的描述。这主要是为了使测验的使用者不至于误用测验和错误地解释测验结果，所以在一般的测验手册中，都有相当的篇幅详细介绍常模团体的大小、取样策略、取样时间以及其他有关情况。这些说明和描述越明确、越详尽越好。

（4）样本大小要适当。所谓“大小适当”并没有明确的指标。根据统计学原理，取样误差与样本大小成反比。所以，在其他条件相同时，样本越大越好。但是还应考虑到人力、物力等方面的因素，通常在决定样本大小时，应注意：①总体的数

目。总体数目小，样本相应可小一些，但不应过小，若总体过小，则可将全部被试入选；当总体较大时，相应的样本也大。②群体的性质，如果群体性质单一，则样本不必太大，即可以反映群体性质；若群体性质复杂，则样本容量(n)就应大一些。③测验结果的精确度。根据统计学原理，抽样误差的大小与样本容量成反比，若要提高精确度，即是说减低抽样误差，就必须加大样本容量(n)。

(5) 常模团体必须是近时的。由于当今教育发展迅速，所以建立的常模必须是近时的，过时的常模是不能作为参照标准的，一个常模不能一劳永逸地使用。例如对瑞文智力测验来说，几年以前所修订的常模对现今可能就不再适用，否则所得智商将产生偏高的趋势。

(6) 注意一般常模与特殊常模的结合。测验手册上所列的常模通常为一般常模，它的适用范围比较广。有时对于某些特殊的群体不一定完全适用。因此，测验在希望使用更为具体的、适合特殊情况的常模。即特殊常模。将特殊常模与一般常模结合起来，可使被试与最接近的群体进行比较。因为各个具体群体在某些方面是独特的，它的成员将与测验手册所列的常模团体成员不符。所以，依据一般常模解释所得的结论可能不够恰当，如果将两者结合使用，解释分数便会更加准确。但特殊常模只提供有关特殊信息，适用范围较窄。所得结论不能在广泛的背景作解释。

二、制定常模的过程

(1) 确定测验将用于哪一个群体。根据测验群体，选定最

基本的统计量，决定抽样误差的允许界限，在此基础上设计具体的抽样方法，并对该群体进行抽样，得到常模团体。

(2) 对常模团体进行施测，并获得团体成员的测验分数及分数分布。

(3) 确定常模分数类型，制作常模分数转换表，即常模量表，同时给出抽取常模团体的书面说明，以及常模分数的解释指南等。

三、几种主要的常模参照分数

(一) 发展量表

人的许多心理特质，如智力、技能等，是随时间而发展的，所以可以将个人的成绩与各种发展水平的人的平均成绩相比较，制定出发展量表。在这种量表中，明确指出个人在按正常途径发展的心理特征处在什么样的发展水平。

1. 心理年龄

比内在本世纪初认为：测量儿童心理成长，可以将一个儿童的行为与各年龄水平的儿童比较，以获得该儿童的心理发展水平。在此设想基础上，他首先寻找并设计出可区分各年龄儿童智力的题目，因为儿童在这些题目上的反应，随着年龄的变化而有系统的改变。每个题目放在大部分的儿童都能成功地完成的那个年龄水平。例如大部分 8 岁儿童都能通过，且有大部分 7 岁儿童不会的题目，就代表 8 岁儿童的智力水平。将该题放在 8 岁的水平内。就每个年龄水平制定适当的题目，可以得到一个可评价儿童智力发展水平的年龄量表。一个儿童在年龄量表上所得的分数，就是最能代表他的智力水平的年龄，这样

的分数就称作智力年龄，简称智龄。所有的年龄量表基本上都是利用相同的推理与步骤制定的，年龄量表将个人的行为与各年龄组的一般儿童比较而给予一个年龄分数。例如，一个儿童能正确回答一般10岁儿童的题目，但对11岁的大部分题目回答不出，则该儿童的智龄为10。

有些测验（如团体智力测验）没有把题目分到各个年龄组。此时，必须首先计算原始分数，即被试在整个测验中正确通过的题数或完成所需的时间，标准比样本中每个年龄组的平均原始分数就作为年龄常模。将被试的原始分数与年龄常模对比，便可求得他（她）的智力年龄。如果某个儿童的原始分数等于8岁的平均分数，则其智龄便是8岁。

因为年龄量表最基本的假设是所测量的特质随年龄作有系统的改变，所以，年龄量表的基本要素是：①一组可区分不同年龄组的题目。②一个常模团体。该团体是由各个年龄的被试所组成的具有代表性的样本。③常模表，即一个表明答对哪些题目或得多少分就该归入哪个年龄的对照表。

年龄常模最大的优点是易于理解与解释，并可以与同年龄团体作直接比较，但必须注意智商的单位不是保持恒定的，而是随着年龄增长而缩小的。例如在3岁和4岁之间的差异，就不等于15岁与16岁之间的差异，因为人在很多方面发展的速率是先快后慢，并随着年龄增长而逐渐减慢，当长到青春期或成年期，便逐渐停止，所以对于这部分被试，年龄常模便不再有任何意义。

2. 年级当量

在教育成就测验中，分数的解释通常也采用年级当量。即将被试的测验成绩与某一年级的学生的平均分数作比较，而说成相当于某一年级水平。这种年级当量选择题目与指定分数的方法与步骤与年龄常模类似，所不同的是用年级水平代替了年

龄水平。例如，一个学生如果能解答六年级的题目或他（她）在测验上的得分与六年级的平均分数相同，则他（她）在该测验上的年级当量便是6。如果标准化样本中，四年级学生在算术测验上正确解答的平均题数是25，则原始分数25的年级当量便是4。

年级常模的单位通常为10个月间隔。在一学年中，假设有两个月的假期，在所测量目标上的发展是不重要的。所以年级当量是5.0，便表示是五年级的初始水平，5.5则表示五年级中期的平均成绩。

年级当量虽然使用普遍，但它也有一些缺点：①教育的内容在各个年级上是不相同的。因此，年级常模只适用于一般课程，而且必须是在各年级间有系统改变，不适合于某些高年级只学1～2年的课程；并且各年级的内容、教学速度都不一样，所以年级单位是不相等的。②年级当量的解释比较困难。例如，一个教育程度较高而又聪明的五年级学生在标准化的教学测验中获得的分数相当于七年级，这并不意味着他已掌握了初一的教学内容，而只是说他在五年级是相当优秀的，并不说明他已具备进入初二的条件；而另一个初二学生获7.9分，则说明他在班中是中等水平，前后两个学生实际掌握的知识并不相等。③年级常模经常被误用为标准。例如，一个六年级的老师就经常希望他班上的全体学生成绩接近或达到六年级常模团体的成绩。这种情况，个别学生可能达到，而大部分学生是不可能达到的。必须清楚，常模与标准是不同的，标准常指所希望达到的标准，常模则是代表群体的次数分布。

（二）商数

过去，曾有许多人企图用两个分数的比率来制订量表，最有名的就是智商。智商最初就被定义为：儿童的智力商数等于

智力年龄与实际年龄的比率。在教育测验中，有时也采用商数来表明教育发展或成就的速率，常见到的有教育商数与成就商数。

1. 教育商数

教育商数（EQ）与智商类似，它是教育年龄（EA）与实际年龄（CA）之比。其公式如下：

$$EQ = \frac{\text{教育年龄}}{\text{实际年龄}} \times 100 = \frac{EA}{CA} \times 100 \tag{7.13}$$

所谓教育年龄是指某岁儿童所取得的平均教育成就。譬如一个学生的教龄为 10 岁，就说明该儿童的教育成就与一般 10 岁儿童教育成就相等。不管年龄的大小，只要测验上所得的分数与某年龄平均分数相等，则教龄便为多少。

教育年龄可以由年级当量间接地得到。例如，一个学生在测验上所得的分数相当于四年级的得分，而四年级学生的众数年龄为 10 岁，则该被试的教龄便是 10 岁。

教育年龄与教育商数和智龄与智商的解释类似，都是表示发展的水平与速率的。但以教龄作单位，有时意义不明确。例如某人的某科教龄是零岁，很可能是他未曾学过，也可能是已忘记掉了。再则有些学科到高年级才开设，而低年级并不开设，此时使用教育年龄作为单位会发生混乱。

2. 成就商数

成就商数（AQ）是将一个学生的教育成就与他智力作比较，即教育年龄与智力年龄（MA）之比：

$$AQ = \frac{\text{教龄}}{\text{智龄}} \times 100 = \frac{\text{教龄/实龄}}{\text{智龄/实龄}} \times 100 = \frac{\text{教育商数（EQ）}}{\text{智力商数（IQ）}} \times 100 \tag{7.14}$$

因为成就商数是将一个学生的教育成就或学业成就与同等智力的学生作比较，所以它不仅可以用来评价学生的努力程

度，也可以用来评价教师的教学效果与质量。前者是因为智力与学业成就二者不等价，智力不够理想，若努力刻苦，仍可获得好的学业成就，此时他的成就商数就较高。反之，成就商数较低，说明该生不够努力，所获得的成就与他的智力不相称。对后者，如果学生的平均教龄低于智龄，说明教学存在问题，未取得应有的效果。

但是使用成就商数来评价学生与教师也存在一些问题。首先是智力与学业成就两者只是中等程度的相关，智力较好，且刻苦努力，并不是就一定能获得好成就。因为学绩测验与智力测验所测量的并不完全是一个东西。其次，到目前为止，任何一种智力测量都不能保证百分之百的可靠，教育测验也同样如此。而使用两个不可靠的分数的比率则更不可靠。虽然有这两点缺陷，但在实际教育工作中成就商数还是有一定的用途的，因为无论如何，低的成就商数是学生与教学不相适应的表现，应该寻找原因，予以补救。当然，在心理测量领域使用得更多的常模参照分数还是百分等级分数和标准分数及其转化形式，这两种分数已在第一节中作过讨论，此处不再重复。

四、呈现常模资料的方法

呈现常模的方法主要有两种：转化表与剖析图。

(一) 转化表

转化表又称常模表，是一种最简单、最基本且最常用的呈现常模资料的方法。它由原始分数、相应的导出分数和对常模团体的有关具体描述 3 个要素构成。有了转化表，使用者便可

以将原始分数转换为导出分数，或从所给的导出分数找到相应的原始分数。

常模表有简单转化表与复杂转化表两种。

1. 简单转化表

简单转化表是将单项测验的原始分数转换为一种或几种导出分数，如表 7.3 所示。

表 7.3　ACT 的百分等级与标准分数

原始分数	百分等级	标准分数
32	99	70
31	96	66
30	89	62
29	78	59
28	67	55
27	54	52
26	42	48
25	31	44
24	21	41
23	13	39
22	6	34
21	1	30
20	1	26

该表是文学院女新生在 ACT 的合成分数（原始分数），百分等级和标准分数（T 分数）的对照表。假若一个学生原始分数为 27 分，则所对应的百分等级为 54，T 分数为 52，分数的意义与解释与本章第一节中的说明完全一致。

利用转化表解释分数时应注意：①只能将分数与表中所描述的常模团体作比较，要和其他常模团体比较，则需其他的常模表。②在没有效度资料时，转化表只能将原始分数转换为另一种分数，而不能作任何推论，即使有效度资料，效标行为也只是从常模资料推论来的。

2. 复杂转化表

复杂的转化表是将包括几个分测验，或几种常模的原始分数与导出分数的对应关系呈现在一张转化表上，如表 7.4 所示。表 7.4 为几个分测验的常模转化表。从表中可以看出，相同的原始分数在不同的分测验上的百分等级不同，而为了得到各分测验上的相同的百分等级，则需要有不同的原始分数。利用此表，可以直接比较一个人在各种分测验上的成绩，但要注意各分测验的资料必须来自同一个常模团体，否则就不能直接比较。

表 7.4　大学生戈登人格问卷的百分等级

男性					女性				
分数	谨慎性	独创性	人际关系	活力	分数	谨填性	独创性	人际关系	活力
38					38			99	
37		99			37		99	98	
36	99	98		99	36	99	98	97	
35	98	97	99	98	35	98	97	96	99
34	97	95	98	97	34	97	96	95	98
33	96	92	97	95	33	96	94	93	97
32	94	89	95	93	32	94	92	91	95
31	91	85	93	90	31	91	89	89	92
30	88	81	90	86	30	87	86	86	89
29	84	76	86	82	29	82	82	82	85
28	79	70	82	77	28	77	77	77	80

27	74	64	77	72	27	72	71	72	75
26	68	58	71	66	26	67	65	66	69
25	62	51	65	59	25	61	58	59	62
24	56	44	58	52	24	55	51	52	56
23	50	37	51	46	23	49	44	45	50
22	44	34	44	40	22	43	37	30	44
21	38	25	38	34	21	37	31	33	38
20	33	20	33	29	20	32	25	27	32
19	29	15	28	24	19	28	20	22	27
18	25	11	23	18	18	24	17	18	22
17	21	8	19	15	17	20	14	15	18
16	18	6	15	12	16	17	12	12	15
15	15	4	12	9	15	14	10	9	12
14	12	3	9	7	14	11	8	7	10
13	10	2	7	5	13	9	6	5	8
12	8	2	5	4	12	7	4	4	6
11	6	1	4	3	11	6	3	3	4
10	5		3	2	10	5	2	2	3
9	4		2	1	9	4	1	2	3
8	3		1		8	3		1	2
7	2				7	2			1
6	1				6	1			

表7.5反映的是不同的被试团体在同一个测验上的原始分数与导出分数的对应关系。

表7.5　不同团体的戈登问卷“谨慎性”分测验常模转换表

原始分数	大学生	中学生	工人	干部
38				99
37			99	98
36	99		98	97
35	98		96	95
34	97	99	93	91

33	96	98	89	86
32	94	97	84	81
31	91	95	78	75
30	88	93	71	69
29	84	90	63	62
28	79	86	55	55
27	74	82	48	47
26	68	78	41	39
25	62	73	34	31
24	56	67	28	24
23	50	61	23	18
22	44	55	18	13
21	38	49	14	9
20	33	42	11	6
19	29	36	9	4
18	25	31	7	3
17	21	26	6	2
16	18	22	5	1
15	15	18	4	
14	12	14	3	
13	10	11	2	
12	8	9	1	
11	6	7		
10	5	5		
9	4	4		
8	3	3		
7	2	2		
6	1	1		

利用这种转化表解释分数，可以提供两方面的信息：一方面它表示出不同团体的导出分数，测验使用者可以将一个人的分数与几个有关常模团体比较；另一方面，它允许对不同团体

作比较。但在解释时必须注意到各个团体的测验分数必须在同样的情况下，即条件一致时获得，否则不便比较。

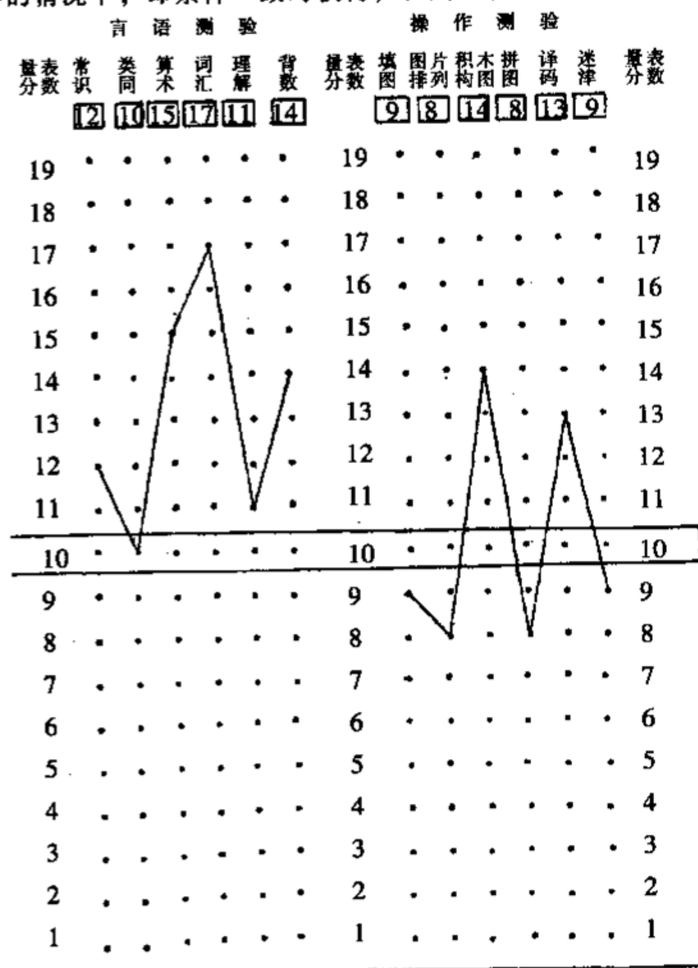


图 7.3　韦氏儿童智力量表剖析图示例

（二）剖析图

剖析图是把一套测验中几个分测验分数同图表（图形）表示出来。从剖析图上可以很直观地看出被试在各个分测验上的表现及其相对的位置。图 7.3 即为一个学生在韦氏儿童智力量表上的剖析图。从图中可以看出，该生总的智商在平均以上，其中言语测验智商较高，操作测验一般，其中词汇、算术背数、积木等较好。

使用剖析图作解释，要求各个分测验所使用的必须是同一个常模团体，否则无法进行比较。

练习与思考

1. 试比较各种导出分数的优缺点。

2. 列举你所了解的各种测验的分数合成方法，并评价它们的合理性。

3. 选择常模团体与制定常模有什么关系？如何选择好常模团体？

4. 离差智商与比率智商的本质差异是什么？

第八章　心理与教育测验的编制与实施

本章提要：

- 编制测验的基本程序
- 测验的实施过程及注意事项
- 解释测验分数的类型与基本原则
- 向受测者报告测验分数的基本原则

任何测量都有测量工具，心理与教育测量的工具通常叫测验（Test）。进行测量的第一步是编制测验，编制出一个好的测验，是实现心理与教育测量科学性的基本前提，同时只有正确使用测验，才能实现一个好的测验的科学功能。因此，本章将讨论在编制和使用测验中的一些基本问题，即编制测验的基本程序与使用测验的若干基本原则。至于编制各种特定测验的具体技术和方法，以及这些测验的实施方法，则在后续章节加以讨论。

第一节 编制心理与教育测验的基本程序

不同性质的心理与教育测验，其编制方法有所不同。学绩测验的编制与能力测验的编制应有所不同，能力测验的编制与人格测验的编制也会有所差异。但不管编制测验的具体技术、过程和方法有多大差异，其基本程序一致。总的来说，编制一个可供使用的标准化的心理与教育测验，一般要经过以下几个步骤：①确定测验目的。②制定编题计划。③编辑测验项目。④预测与分析。⑤合成测验。⑥测验标准化。⑦鉴定测验。⑧编写测验说明书。下面分别进行简要讨论。

一、确定测验目的

在这一步主要要解决3个问题：

（一）明确测量对象

明确测量对象，也就是明确测量哪些个人或团体。通常以年龄、性别、职业、受教育程度、经济状况、民族、文化背景等指标来区分测量对象。施用于不同对象的测验应该有其不同的特征，而不应千篇一律。

（二）明确测量目标

明确测量目标，也就是明确测量什么心理功能，是测能力、人格，还是测学业成绩。不仅如此，还要进一步把目标具体化。例如，若要测量人的态度，必须按照态度的定义分为认知方式、情感表达和行为倾向 3 个层面，并给出这 3 个层面的操作性定义，然后按照操作性定义编制测题。若要测量智力（一般能力），就必须把智力分解为若干因素，并具体规定各种因素的意义。如美国心理学家瑟斯顿通过因素分析，将智力分解为 7 个基本因素：

（1）语文理解——阅读时了解文义的能力。

（2）词语流畅——准确迅速拼词与敏捷联想词义的能力。

（3）数学运算——准确迅速运用数字解答数学问题的能力。

（4）空间关系——准确迅速判断空间方向与空间位置关系的能力。

（5）机械记忆——对事物进行强记的能力。

（6）知觉速度——准确迅速观察和识别事物的能力。

（7）一般推理——根据已知判断推出未知判断的能力。

瑟斯顿根据上述定义和分析，编制了《基本心理能力测验》（1941）。把目标具体化，是保证测量可靠性的基本条件。

（三）明确测量用途

明确测量用途，也就是明确编制的测验干什么用，是用于描述受测者的心理特质，还是用于诊断心理是否异常，是用于选拔人员，还是用于验证某个理论假设。用途不同，编制测验时的取材范围以及测题的难度也不尽相同。

二、制定编题计划

编题计划是编制测验的总体构思。编题计划要明确的信息主要有两个方面：一是全面而具代表性的测验内容，不致使测题偏离了应测的范围；二是对各个内容点的相对重视程度，通常用百分比来标明。

编题计划主要有两个用途：①编题计划指明了应该编哪些方面的测验项目以及编多少个项目，因此，测题编制结束后，可比照计划核对测验项目是否反映了所要测量的领域。②在记分时可按计划中百分比确定每类测验项目的分数标准。

三、编辑测验项目

在编辑测验项目时需要解决 3 个问题：

(一) 收集测验资料

一个测验是否有效，取决于该测验是否能够测得研究者所要测得的东西，为此，就需要收集适当的测验资料。尽管不同性质的测验所依据的资料内容各异，但都必须遵循几个共同的原则：

(1) 资料要丰富。资料收集愈齐全，编题工作愈顺利。无论是能力或人格，均是十分复杂的复合性心理结构，不能仅凭一两种简单的项目去推断一个人的智愚或人格特征，必须包含许多不同类型的材料。例如，编制人格测验，需要收集描述人格特征的大量词汇、临床观察的资料、已有的人格测验量表中的测题等。

(2) 资料要有普遍性。这有两层意思：一是当编制智力类测验时，所收集的资料对于不同文化背景、不同经济地位、不同地区的个人或团体应当是公平的，应尽可能避免特殊知识经验对测验结果的影响；二是当编制人格测验时，所收集的资料应当能够全面反映某一文化背景中的团体的基本人格特征。

(3) 资料要有趣味性。资料的趣味性可以减少受测者由于缺乏足够的动机而引起的测量误差。

(二) 选择项目形式

在心理测量中，必须将测验项目以某种形式呈现给受测者，而测验项目呈现的形式又取决于受测者的年龄、人数的多少、测量的目的、测验项目的性质等因素。因此，在选择测验项目形式时，应当注意将这些因素考虑进去。例如，在学绩测验中，如果要考察对概念和原理的记忆，宜用简答题；要考察综合运用知识的能力则宜用论文题。再如，在智力测验中，对于幼儿、文盲或识字不多的人，宜用口头测验；对聋哑人，则

宜用操作测验；受测人数过多，且时间、财力有限时，宜用团体测验，而在受测人数较少时，可用个别测验。

对于测验项目的确定，我国心理学家廖世承、陈鹤琴早在几十年前曾提出的几条原则现在仍可供参考：①使受测者容易明了测验方法。②使受测者在完成测验时不会因测验项目的形式不当而作错。③测验过程省时。④计分省时省力。⑤经济。

（三）编写测验项目

编写测验项目是一个反复的过程。在这过程中，测验项目编制者需要对测验项目进行反复修改，其中包括订正意思不明确的词语，删改一些重复和不适当的项目，增加有用的题目等等。

在编写测验项目时要注意：

（1）测验项目的取样应当对欲测心理品质具有代表性。只有测验项目真实反映测量对象的特征时，才能保证测验结果的有效性。

（2）测验项目的取材范围要同编题计划所列项目范围相一致。

（3）测验项目的难度应有一定的分布范围。如果是能力测验或学业成就测验，就应当包括各种不同难度的测验项目，以鉴别各种不同能力或不同知识水平的人员；如果是人格测验，就应当选编那些在不同方向的备选答案上都有一定人数分布的项目，以鉴别具有不同人格特征的人员。

（4）编写测验项目的用语要力求精炼简短，浅显明了。

（5）初编题目的数量要多于最终所需要的数量，以便筛选或编制复本。

（6）测验项目的说明必须简明。

四、预测与项目分析

初编的测验项目是否具有适当的难度和区分度，必须通过预测进行测验项目分析，以便进一步修改。

（一）预测

预测的目的在于获得被试对测验项目做何反应的资料。它既能提供那些题目意义不清、容易引起误解等质的信息，又能提供测验项目优劣的量的指标。

预测应注意的问题是：

(1) 预测对象应取自将来正式测验时准备施用的群体，虽然人数不必太多，但要具有代表性。

(2) 预测的情境应力求同正式测验的情境一致。

(3) 预测的时限可以适当延长，以便每一受测者都能将题目做完。

(4) 施测者应对受测者的反应加以记录。如在同一时限内，受测者所完成的题数，以及受测者反应的题意不清之处等，以便修改项目时参考。

（二）项目分析

测验项目分析就是对预测结果进行统计分析，确定项目的难度和区分度。由于预测的受测者样本小可能会存在取样误差，由此获得的项目分析结果未必完全可靠。所以，需要对来自同一总体的两个样本施测，然后分别进行测验项目分析，看对两个样本的分析结果是否一致。关于项目分析的具体原理与

技术问题，请参见第六章。

五、合成测验

合成测验就是把经过预测以后证明有价值的项目排成有组织的测验。它要解决的问题是两个：一是测验项目的选择；二是测验项目的编排。如要编制复本，还须懂得怎样编制复本。

（一）测验项目的选择

选择测验项目的指标有三：一是测验的性质，即要选择那些能够测量所要测量的东西的项目。假若要测量的是语言推理能力，就不能选择测量阅读能力或运算能力的项目。二是项目的难度。选择多大难度的项目并无固定的标准，选拔性测验要求难度大些，考察性测验则要求难度不可太高，人格测验则不要求难度。三是项目的区分度。一般来说，项目的区分度越高越好，对于选拔性测验尤为如此，但有时也可以保留若干区分度不高的项目，这要视项目的重要性而定。

（二）测验项目的编排

测验项目选出之后，需要加以合理安排。在测验开头应该有一、两个较容易的项目，以使受测者熟悉作答程序，解除紧张情绪，建立信心，较快进入测量情境。对测验项目的总的编排原则是要由易到难，这样可以避免受测者在难题上耽搁时间过多，而影响对后面问题的解答。在测验最后可有少数难度较大的题目，以测出受测者的最高水平。

下面是两种常见的测验项目的排列方式：

（1）并列直进式：此种方式是将整个测验按测验项目材料的性质归为若干分测验，对于同一分测验的测验项目，则依其难度由易到难排列。

（2）混合螺旋式：此种方式是先将各类测验项目依难度分成若干不同的层次，再将不同性质的测验项目予以组合，作交叉式的排列，其难度则渐次升进。此种排列的优点是，受测者对各类测验项目循序作答，从而维持作答的兴趣。

（三）编制复本

为增加实际的效用，一种测验至少要有等值的两份，份数越多，使用起来越便利。所谓等值需要符合下列条件：

（1）各份测验测量的是同一种心理特质。

（2）各份测验具有相同的内容和形式。

（3）各份测验不应有重复的项目。

（4）各份测验项目数量相等，并且有大体相同的难度和区分度。

只要有足够数量的测验项目，编制复本的手续是很简单的，先将所有可用的项目按难度排列，其次序为1、2、3、4、5、6…… 如果要分成两个等值的测验本，可采用下面的分法：

A本：1、4、5、8、9、12、13、16、17、20……

B本：2、3、6、7、10、11、14、15、18、19……

如果要分成3个等值的测验本，可采用下面的分法：

A本：1、6、7、12、13、18、19、24……

B本：2、5、8、11、14、17、20、23……

C本：3、4、9、10、15、16、21、22……

采用上面的分法可使各复本之间在难度上基本相等，从而获得大体相同的分数分布。复本编好后，应该再测验一次，以

确定各份测验究竟是否等值。

六、测验标准化

一个测验的好坏，取决于对该测验的标准化水平。所谓标准化是指测验的编制、施测、评分以及解释测验分数的程度的一致性。具体地说，测验标准化包括下列内容：

（一）测验内容

标准化的首要前提，是对所有受测者施测相同的或等值的题目，测验内容不同，所测得的结果无法比较。

（二）施测过程

标准化的第二个条件是所有受测者必须在相同的条件下施测。其中包括：

（1）相同的测验情境：如统一的采光条件，统一的桌椅高度，统一的桌面面积，统一的场所布置等。

（2）相同的指导语：指导语一般包括两部分，一是向受测者说明测验的目的，以便解除受测者的顾虑；二是向受测者说明如何对测验项目反应。指导语必须事先拟好，印在测验项目的前面，并且力求清晰、简单、明了，不致引起误解。对受测者不熟悉的测题类型，应当有一至二个例题。

（3）相同的测验时限：测验的时间限制是测验程序中的重要方面。不过，不同的测验对时限要求很不相同。一般来说，人格测验对时限的要求不太严格，甚至不要求时间限制；但能力测验和学绩测验必须考虑时限问题。确定时限一般采用尝试

法，即通过预测来决定。通常的时限定为大约 90%的受测者在预定的时间完成全部测验项目即可。

（三）测验评分

评分的客观性是标准化测验的第三个条件，评分的客观性意味着两个或两个以上的评分者对同一份测验试卷的评定是一致的。只有当评分是客观的时候才能将分数的差异归于受测者本身的差异。但要做到完全客观（一致）的评分是较困难的。一般来说，不同评分者之间的一致性达到 90%以上，便可认为评分是客观的。客观性评分要求：

（1）对反应要及时清楚地记录，以免由于记忆丧失造成混乱，尤其是在口头测验和操作测验中更应如此。

（2）要有一张标准答案或正确反应的表格，即记分键。选择题测验的记分键包括每一测验项目正确反应的号码或字母；问答题的记分键包括一系列的正确答案和允许变化的范围；论文题的记分键包括一致可接受答案的要点；人格测验没有正确答案，记分键上指明的是具有或缺少某种人格特征者的典型反应。

（3）将受测者的反应与记分键比较，确定受测者反应应得的分数。

（四）测验分数的解释

一个标准化的测验，不仅指测验内容、施测过程和评分程序的标准化，而且指对测验结果的解释的标准化。如果对同一测验结果（分数）可做出不同的解释，那么测验便失去了客观性。

某一测验分数只有与一定的参照标准相比较，才能显现出它所代表的意义。在心理测验中，建立参照标准的过程也就是

建立常模的过程。建立常模的方法，请参见第七章。

七、鉴定测验

测验编好后，必须对其可靠性和有效性进行鉴定，以便确定该测验是否可用。对测验的鉴定，主要是确定其信度系数和效度系数。

（一）信度（reliability）

信度指的是测验的可靠性，即用同一测验多次测量同一团体，所得测验结果之间具有一致性。我们用钢片卷尺去测量一木杆的长度，所得结果是可靠的，因为无论是由一个人数次测量，还是分别由数个人去测量，所测得结果都是一致的。如果改用橡皮软尺去测量木杆，一人多次或多人测量结果就难以一致，也就是说，这一测量工具是缺乏信度的。由此可见，信度是衡量测验质量的最基本的指标，因而测验编好后首先要鉴定该测验的信度。

（二）效度（validity）

效度指的是测验的有效性，即一个测验在多大程度上能够测得它所要测得的东西。如果一个测验的效度很低，那么说明该测验所测得的东西不是它所要测的东西。例如，智力测验所要测的东西应是智力，假如它测得的是知识或人格，那么就说明这个智力测验对于测量智力是无效。因此，测验编好后，还必须检验该测验的效度。

（三）测验量表与常模

任何测量都是以数量化的形式来表达测量结果的。心理测量是以心理测验为测量工具的，它必须采用一定的量表作为标准化的记分制度，来实现测验结果的数量化。所以，测验编制者为了说明和解释测验结果，必须根据测验的性质、用途以及所要达到的测量量表的水平，按照统计学的原理，把某一标准化样本的测验分数转化为具有一定参照点、等值单位的导出分数，这就是所谓的测量量表。在心理测验中，常见的测验量表有百分等级量表、标准分数量表、T 量表、发展量表、智力商数量表等。如果将标准化样本的测验分数与相应的某一或几个测验量表分数一起用表格的形式呈现出来，就是该测验的常模表。标准化的心理测验都在测验手册中提供有可供解释测验分数的常模表。

八、编写测验说明书

测验说明书向测验使用者说明如何使用该测验，以此才能保证测验的信度和效度。说明书应包括下列内容：

（1）本测验的目的与功用。

（2）本测验的理论依据。

（3）测验内容及实施测验的方法，包括①何种测验；②内容分几部分；③每部分有多少测验项目；④如何作答等等。

（4）测验的标准答案和评分方法。

（5）关于测验的信度、效度资料的说明。

（6）常模表，即如何依据常模解释测验结果。

第二节　测验的实施

一个经过信度和效度检验证明可用，并已获得常模资料的测验便可正式出版交付使用了。测验的使用主要涉及两个问题：①如何实施测验才能保证测验分数尽可能少受施测过程的影响；②如何解释测验分数才能保证受测者的心理不受负面影响。本节围绕着上述两个问题讨论使用心理测验的一些技术性问题。

一、测验的实施过程

心理与教育测验的基本原理在于，通过观测受测者在测验情境中的行为样本，可以推断他平日的一般行为特征。换句话说，根据测验分数，可以预测受测者可能会产生什么心理症状，或可能做出哪方面的工作成绩等等。但实际测验分数不仅受到与测验目的有关的变量的影响，而且也可能受到与测验目的无关的变量引起的影响。换言之，测验分数不仅决定于测量工具本身，也受测验过程的影响。因此，在施测过程中，施测者应当了解哪些因素会影响测验分数，并进一步对这些因素进行有效的控制。

（一）施测前的准备工作

一个好的施测程序中最重要的就是预先做好准备。

（1）准备好测验材料。施测者必须把施测中所要用的材料按一定顺序放置在适当的位置，使受测者易于看到和找到。例如在操作测验里，要求受测者拼一个马图，施测者必须将马图的几个碎片按规定的顺序和位置放在受测者面前。如果不事先熟记放置的顺序，到时必定会手忙脚乱，安放不合规定，以致有的受测者可能因碎片的位置引起对马的某一部分的联想使他易于得分，而另一些受测者可能相反，丢失了不应丢失的分数。大部分智力测验表都有操作测验，操作材料的放置都有相应的规定，因此，都必须事先做好准备。

（2）熟练掌握施测手续。为了掌握施测手续，必须对施测者进行必要的训练。训练的内容包括：①熟悉测验内容；②掌握施测步骤；③掌握记分方法；④掌握解释分数的技术。

（3）熟记测验指导语并能用口语清楚而流利地说出来。凡是要求念读的指导语都不应念错、停顿、重复或结结巴巴，否则会影响测验分数。

（二）指导语

心理测验的指导语通常包括对测验目的的说明和对题目反应方式的解释。指导语直接影响受测者反应的态度和方式。有人曾以三种不同的指导语对三组受测者实施同一智力测验，结果是将该测验说成“智力测验”的一组成绩最好，而将它说成“日常测验”的一组成绩最差。人格测验常涉及一些受测者敏感的问题，因此，指导语不适当，会造成很多不利的影响。

指导语的主要作用是使受试者按正确形式对题目作出反应。确定指导语时，要注意不要暗示受测者应该选择什么样的

答案。当指导语中包括对测验目的的说明时，更应注意这一点，一般要求测验的主持者和指导语都应保持和表述中立的态度，不倾向于答案中的任何一种方向。

一般的能力测验和成就测验都要求有标准严格的时间限制，因为速度是能力测量中的一个重要因素，而人格测验和态度测验一般不要求有时间限制。

（三）测验情境

测验情境包括测验场地（通风、光线、噪音）、座位、答案纸型等。这些因素都会影响测验分数，因而需要加以必要的控制，使之对每一个受测者都保持相同。标准化测验一般都对测验条件做出严格的规定，其中包括采光条件、桌椅的高度、桌面的面积，测验采用的试卷都用同一种纸张按同一规格印刷，受测者答题时所用的铅笔一般由施测者统一提供等。

这里尤其要强调的是，心理测验进行之时，务必不能有外界干扰。为此，测验室的房门应挂上一个牌子，示意测验正在进行，旁人不许进入。团体测验时，可以把屋门锁上或派一名助手在门外守候，阻止晚来者入场。

施测者的状态对测验分数也有影响，施测者的语言、行为、态度、表情等都要严格控制。

（四）测验焦虑

测验焦虑是指被试因接受测验而产生的一种忧虑和紧张情绪，它会影响测验结果的真实性。例如进行操作性测验时，由于过度紧张会使手眼失去良好的协调；又如考试之前要求学生定出得分指标为 90 分，有一两题做不出（每题 10 分），一个上进心很强的学生就会产生不安情绪。因此，在测验时，应注意稳定被试的情绪。主试有时可以利用保证测验结果绝对保密

或鼓励被试等方法来消除测验焦虑。心理学的有关研究证实：①能力与测验焦虑成负相关，亦即能力愈高的人，测验焦虑愈低。②抱负水平与焦虑成正相关，也即愈渴望得高分，测验焦虑愈高。③竞争性测验的测验焦虑高，经常接受测验的人焦虑低一些。④轻微的测验焦虑会增进测验效果，但焦虑太高或毫无测验焦虑，则会降低测验效果。

实施测验时，主试的以下 4 种态度容易使被试产生过度的焦虑，应尽可能避免。

(1) 以测验来威胁被试，以使被试循规蹈矩。

(2) 警告被试一定要尽力，因为"这项测验很重要"。

(3) 告诉被试答题要快，才能在规定的时间内答完。

(4) 恐吓被试说："如果测验失败，会有严重的不良后果。"

(五) 与受测者建立良好的协调关系

在心理测量学中，良好的协调关系指的是施测者努力设法引起受测者对测验的兴趣，取得他的合作，以保证他能按照标准测验指导语行事。在做能力测验时，应要求受测者认真集中注意于当前的任务，并要求他尽最大的努力来完成它；在填写人格问卷表时，应要求他坦率而忠实地回答问题；在做投射性测验时，则要求他将由刺激唤起的联想充分报告出来，如此等等。总之，施测者要力图激起受测者尽量地并有意识地按照指导语去做。

根据测验性质的不同、受测者的年龄以及其他特点的不同，建立良好协调关系的技巧也有所不同。在测试学龄前儿童时，就要考虑到儿童对陌生人羞怯，注意力分散等特点。施测者以友好、愉快、放松的自然态度可以使儿童感到信任，那些害羞、胆小的儿童需要较多的时间来熟悉情境。因此，施测者

不能操之过急，匆忙示范，耐心等待儿童到他愿意接触时再开始。测验要像玩游戏一样呈现给儿童，幼儿有时会拒绝测验，有时没有兴趣，测验手续就要相对灵活一些。对小学一二年级甚至三年级的小学生，测验也要像做游戏似的才容易引起他们的兴趣。再大一些的学生则通过竞赛精神去激发他们做好测验。

在测验学校儿童或成人时必须记住，每个测验都暗含有暴露某个人弱点的可能。例如，这个题目答不出来，那个拼图不成功，这都会使人感到丢面子。因此，测验一开始就可以说清楚，没有人能够正确答出所有这些题目。这样交代几句是有好处的，否则他们在遇到困难题目时就会体验一种失败的挫折感，甚至影响到不能在规定时限内完成其它测验。

鼓励受测者努力完成测验，争取他的合作，使他表现出真实水平或实际情况，这并不是说在受测者不会做时可以给他提示、暗示或者任何方式的帮助，这样做同样会使测验分数失去作用。

（六）评分技术

在标准化的心理测验中，测验与答卷通常是分开的。被试将测验项目的答案直接记录在专用答卷上。另外备有一份标准答案卡，此为记分键，评分时只要将被试的答案逐一与标准答案相比较，即可评定被试应得的分数。有时候为了节省评分的时间，采用记分板来记分。所谓记分板是把一张空白答案纸上的正确答案打成圆形或方形的洞，评分时只要将记分板套在每一张答案纸上，然后统计从洞中出现的正确答案之数目即可。凡洞中未出现任何记号者，需以红笔画上斜线，这样可让被试知道答错了哪些题。

二、测验分数的解释

测验分数的解释涉及两个问题：一是如何看待测验分数的意义；二是如何将测验分数的意义告诉给受测者。

（一）如何看待测验分数的意义

施测者进行一个测验结果的解释必须一方面对所做的具体测验（包括它的常模的代表性、信度、效度、难度等）要熟悉了解，另一方面对受测者的情况（文化程度、职业、是否可能接触测验中的有关问题等）也要有所了解。此外还必须结合当时测验的具体情况，例如是否有干扰，受测者当时有无情绪波动或身体不适等综合考虑。同一个分数可能是由于不同原因造成的，合格的施测者会结合以上三方面的因素对测验分数作解释，对同一分数可作出不同解释。例如，用平均初中文化程度的标准化样本的智力测验来测量一个不够初小文化程度的受测者，如果测得 IQ 为 85，就可以认为他基本上是中等智力水平；如果受测者原来文化程度是大学毕业，也测得 IQ 为 85，就可解释为受测者可能因疾病而使智力有所减退，属于中下水平。

关于测验分数的解释，高德曼（Goldman）曾提出一个含有三个维度的解释模型，可作为解释分数的参考。这三个维度分别是解释测验分数的类型、资料处理的方法和资料的来源。他提出解释测验分数的 4 种类型：叙述的解释、溯因的解释、预测的解释及评价的解释。资料处理的方法有两种：机械的处理与非机械的处理。资料的来源有两种：测验资料与非测验资

料。将此三个维度加以组合，可有 $4\times2\times2=16$ 种不同的解释方式。

就资料的来源而言，有测验资料和非测验资料。前者系指由各种标准化测验所得到的分数；后者则包括学校成绩、家庭背景、晤谈或观察所得资料。

就资料的处理方法而言，有机械的处理与非机械的处理。前者又可称为统计的处理，包括常模对照表、预期表、侧面图分析及回归预测等；后者又可称为临床诊断的处理，采用归纳与演绎的推理方法，综合评判资料的意义，此种方法比较主观、直觉与不清楚。

就解释的类型而言，上述 4 种解释类型代表了 4 种不同层次的解释方式。每种解释类型的含义如下：

(1) 叙述的解释：指描述个人的心理特征状态。例如，这个学生是一位怎样的学生？聪明的？中等的？或愚笨的？他的语文推理是否优于非语文推理？他喜欢做些什么？有什么样的性格特点？

(2) 溯因的解释：指追溯过去以解释个人目前的发展情况。例如，他为什么会这样？他的阅读困难是否是情绪困扰的结果？或缺乏基本的阅读技能？或缺乏学习的兴趣？他拒绝机械的学习活动是否由于父母的压力？或过去的失败？或兴趣太广泛所致？

(3) 预测的解释：指推估个人未来的可能发展情形。例如，他上高中的成绩会怎样？他升入大学的可能性有多大？他在理科方面的发展是否比在文科方面的发展更能成功？他是否可能成为一个问题青年？

(4) 评价的解释：指作价值的判断或做决定。此种解释是依据上述几种解释而作的判断。例如，准许人高中或大学、雇用人员、编班等均是属于此种解释。他应该学习什么样课程？

进什么样大学？他应该成为工程师或商务经理？

在解释测验分数的意义时，应遵循以下几个基本原则：

(1) 主试应充分了解测验的性质与功能。测验使用者必须具备心理测验的基本知识与概念，方能了解测验的性质与限制。任何一个测验都有其编制的特定目的和独特的功能，使用者在解释之前必须从其编制手册中，详细了解编制过程的标准化及测验的信度、效度、常模等是否适当。更重要的，应知道测验能测量什么，不能测量什么，分数在使用上有何限制。有时两个测验的类型虽然相同，但测量的功能往往不同。例如，韦克斯勒智力量表和瑞文标准推理测验都是智力测验，但内部结构有很大的不同，所能发挥的作用也有区别。再如，卡特尔16PF测验与明尼苏达多相人格调查表都是人格测验，但后者更多地发挥临床诊断的功能，前者则更多地针对正常人。在教育测验里，也是这样。同是算术测验，有的偏重简单的计算技能，有的却偏重推理能力；同是科学能力测验，有的注重测量科学术语的基本知识，有的却注重测量科学原理的应用。对以上这些有了正确的认识，方能作客观的解释。

(2) 对导致测验结果的原因的解释应慎重，谨防片面极端。一个人在任何一个测验上的分数，都是他的遗传特征、测验前的学习与经验，以及测验情境的函数，这3个方面对测验成绩都有影响。所以我们应该把测验分数看成对受测者目前状况的测量，至于他是如何达到这一状况的，则受许多因素影响。

为了能对分数作出有意义的解释，必须将个人在测验前的经历或背景因素考虑在内。比如，在词汇上得到相同的分数，对于大城市的孩子与边远山区的孩子具有不同的意义。惠勒曾于1932年测量了美国某山区儿童的智力，发现6岁以前的儿童，其智力与常模相近；6岁以后与常模的差距随年龄递增，

这是由于环境影响，得不到平等学习机会的结果。

测验情境也是一个需要考虑的因素。比如，一个学生可能因为身体不适，情绪不佳，不明了施测者的说明或受到意外干扰，这些都会产生测验焦虑。如果对这些因素控制得不好，就会使分数受到影响。在这种情况下，应当找出造成分数反常的原因，而不要单纯以分数武断地下结论。

(3) 必须充分估计测验的常模和效度的局限性。为了对测验分数作出确切的解释，只有常模资料是不够的，还必须有效度资料。没有效度证据的常模资料，只告诉我们一个人在一个常模团体中的相对等级，不能做预测或更多的解释。在解释分数时人们最常犯的错误就是仅根据测验的标题和常模数据去推论测验分数的意义，而忽略效度的不足或缺乏。假若一个测验的名称是内外向量表，并有可利用的常模资料，那么就很容易把得高分的人说成是内向性格，即把它当作有效度资料那样来解释。

即使有了效度资料，在对测验分数作解释时也要十分谨慎。因为测验效度的概化能力是有限的，不同的常模团体和不同的施测条件，往往会得到不同的结果。在解释分数时，一定要依据从最相近的团体，最相匹配的情境中获得的资料。

(4) 解释分数应参考其他有关资料。测验分数不是了解学生的唯一资料，为正确了解其心理特质尚需参考其他有关资料。只凭学生的单一测验分数解释其心理状态，容易作出错误的解释。例如，某生在智力测验上得到 IQ 为 80，在不考虑其他资料的情况下，只能解释："甲生的智力属于中等偏下。"但是，如考虑他的在校成绩时，解释可能大不相同。如果他的在校成绩经常保持在年级前五名，则不可能作出如上的解释，可能需要进一步探讨他在做测验时的动机、态度、情绪与健康状况等。有了这些资料作为佐证，才能正确判断其智力是否全部

正常发挥，测验结果是否可靠。

同样的，解释时亦须参考其它的测验资料，只凭单一的测验分数加以解释，也可能全然不同于综合考虑几个测验分数。例如，根据自陈量表测验的分数，某生的性压抑分数高于平均数两个标准差；但在投射测验中有关性的反应，却高于平均数一个标准差。如仅依自陈量表的分数解释时，只能解释说："某生的性压抑倾向甚强。"但如果参照投射测验的分数综合解释时，则可解释说："某生的性兴趣强于一般人（投射测验），但他却将性兴趣加以严重的压抑（自陈量表）。"

总之，测验分数的解释应尽可能参考其他的资料，如教育经验、文化背景、面谈内容、习惯、态度、兴趣、动机、健康、语文程度及其他测验的资料。唯有如此，解释才能更客观且更深入。

(5) 对测验分数应以"一段分数"来解释，而不应以"特定的数值"来解释。由于每一个测验均会受到测量误差的影响，因此在解释测验分数时也应考虑到测量误差的存在。测量误差的大小与信度的高低有关。信度越高，则误差越小。但永远不可能完全消除误差，因此，应该永远把测验分数视为一个范围而不是一些确定的点，也就是要对测验分数提供带状的解释。倘若使用确切的分数，应说明这些分数不是精确的指标，而是我们对某人真实分数的大体估计。

(6) 对来自不同测验的分数不能直接加以比较。即使两个测验名称相同，由于所包含的具体内容不同（因而所测量的特质不完全相同），建立标准化样本的组成不同，量表的单位（如标准差）不同，其分数也不具备可比性。如来自两个智力测验的分数，在没有其他信息的情况下，我们无法判断谁高谁低。

为了使不同测验分数可以比较，必须将二者放在统一的量

表上。当两种测验取样于相同范围时，人们常用等值百分位法将两种测验分数等值化。具体做法是：将两个测验都对同一个样本进行施测，并把两种测验的原始分数都转换成百分等级，然后用该百分等级作为中转点，就可以做出一个等价的原始分数表。如果某人在测验 A 中原始分数 55 是 90 百分等级，而测验 B 中原始分数 36 也是 90 百分等级，那么他在测验 A 获得的 55 分就与在测验 B 获得的 36 分等值。(详见第九章)

(二) 如何向受测者报告测验分数

如何向当事人及与当事人有关的人员（如家长、教师、雇主等）报告测验分数，使他们更好地理解分数的意义是一件非常重要的事。下面所列举的一些原则，可供报告测验分数时作参考。

(1) 使用当事人所理解的语言。测验像其他特殊领域一样，具有自己的词汇，因此你所理解的词并不意味着当事人也一定理解。例如，你懂得标准差和标准分数，然而当事人可能不懂。因此你必须用非技术性的用语来解释标准分数，可以把它解释成相对位置（即百分等级）。必要时可以问问当事人是否听懂，让他说说你的解释是什么意思。

(2) 要保证当事人知道这个测验测量或预测什么，这里并不需要作详细的技术性解释。例如你并不需要向当事人解释职业兴趣，并将他与从事各种职业的人加以比较，如果在某一方面得了高分，就意味着如果他参加这个工作可以长期干下去。但也不能过于简单，只告诉当事人某个量表的题目或测量什么是不够的，这在具有情绪色彩的人格特征测量方面特别重要。例如，对人格测验中的男性化、女性化量表就要加以解释，以免受测者误解。

(3) 如果分数是以常模为参考的，就要使当事人知道他是

和什么团体在进行比较。例如，同一个百分等级对于普通学校和重点学校其意义是不同的。

(4) 要使当事人认识到分数只是一个估计。由于测验的信度、效度不足，分数可能有误差，而且对于一个团体总体来说有效的测验，不一定对每个人都同样有效，但也不能让受测者感到分数是毫不足信的。

(5) 要使当事人知道如何运用他的分数。当测验用于人员选拔和安置问题时这点是特别重要的。要向当事人讲清测验分数在作决定过程中起什么作用，是完全由分数决定取舍，还是只把分数作为参考；有没有规定最低分数线；测验上的低分数能否由其他方面补偿等等。

(6) 要考虑测验分数将给受测者带来什么影响。由于对分数的解释会影响受测者的自我认识、自我体验和自我评价，所以在解释分数时要把对分数意义的解释和必要的咨询工作结合起来，以免受测者因分数不理想而造成自卑心理。

(7) 测验结果应向无关的人员保密。当事人的测验分数不应让其他无关的人员知道，以免对当事人造成不良的影响。因此，分数的报告采用个人的解释为宜，不宜采用团体解释或公告通知的方式行之。

(8) 对低分者的解释应谨慎小心。在测验上获得低分数者或分数不理想者易有自卑或自我贬抑的心理产生。因此，对这些当事人报告测验分数时，态度要诚恳，措词要委婉，避免作直接了当的解释。例如，智力测验得到 IQ65 者，勿作这样的解释："你属于智力缺陷者。"较理想的解释应是："这个分数表示你的学习能力比一般人低了一点，但是有些像你这种能力的人，由于刻苦努力而有很不错的表现。"

(9) 报告测验分数时应设法了解当事人的心理感受，并采取适当的措施加以引导。报告测验分数时，宜先让当事人充分

表达测验时的心理感受，如他的动机、态度、情绪、注意、健康等，以便知道他的测验分数是否代表在最佳的情况下所作的反应。例如，某学生表示他在做智力测验时情绪很恶劣、心不在焉；而另一位则表示他在做测验时，动机强烈、注意力集中。虽然两位学生得到相同的IQ为115，但代表的意义可能迥然不同。

同样的，解释完分数后宜鼓励当事人表达对测验结果的感受，如发现当事人对分数有误解或不良态度，应立即配以咨询，予以适当的引导，以免给当事人造成自卑心理或其它不良影响。

练习与思考

1. 结合实例简述编制一个心理测验的基本程序。
2. 阐述实施心理测验应注意的问题。
3. 如何正确解释测验分数的意义。
4. 论述向当事人报告测验分数的基本原则。

第九章　测验等值

本章提要：

- 测验等值的实质
- 测验等值的条件
- 测验等值的基本计算方法
- 常用等值设计
- 测验等值误差估计

第一节　测验等值概述

一、测验等值来源于测量实践的需要

在心理与教育测量实践中，经常遇到一个测验需要配备多个测验形式的情况，特别是那些测验内容易受记忆或易受针对性训练影响的测验，在测验之前需严格保密，测验之后不能再用，必须配备多个不同形式供不同次施测所用。对于这种情况，测验编制者显然希望这些不同形式应该是“相等”的，也就是说，如果是对同一个被试，各个不同形式所测结果应该是完全一样的。为达此目的，测验编制者做了许多努力，但在实际施测后，不同形式之间的差异依然存在，这就会引起评价的不公。这种现象在需要对参加不同形式施测的被试作统一评价时就会造成一些明显的失误。比如我国的高等教育自学考试，同一专业，同一学科年年施测，对考生的评价主要是区分及格还是不及格，其形式就是上不上 60 分。如果各年所用形式之间不相等，就可能出现许多误判，或者把及格学生判成不及格，或者把不及格学生判成了及格。这对考生来说是不公正，对社会来说是无信誉。避免这种失误的一条途径是寻找到不同测验形式之间分数的转换关系，把所有不同形式测验的分数都转换到同一个分数系统上，就不再会出现上述不公正现象了。测量学上把为达到这一目的而发展起来的一套专门技术称为测验等值（Test Equating）。测验等值在国外已经有了许多成功

的应用，国内，在诸如高考这一类大规模正式考试中的研究与应用也已起步。

二、测验等值的实质

从本质上来说，测验等值就是通过对考核同一种心理品质的多个测验形式作出测量分数系统的转换，进而使得这些不同测验形式的测验分数之间具有可比性。在实际操作中，测验等值可使各个不同形式的测验分数均对应起来，测验主持者可以任意指定其中的一个分数形式作为基准，而使所有其它形式的分数都转化到这个基准形式上。比如，经过等值计算，B测验形式的85分对应于A测验形式的82分，C测验形式的80分也对应A测验形式的82分，A、B、C三种形式施测结果均可以A形式分数报告，即参加A测验形式得82分，参加B测验形式得85分，参加C测验形式得80分的3个被试，均可报告他们在该测验上得分为82。因为所测3个被试的水平是一样的，而在不同形式上的施测分数的差异，仅是由于命题难度把握不稳引起的表现形式差异。也可以认为，如果一个被试在A测验形式上得分82，则参加B测验形式他将得85分，参加C形式他将得80分。

测验等值中所说的测量分数系统的转换与测验原始分数与导出分数之间的转换是不相同的。等值转换的目的是为了比较两个不同测验形式之间的实测分数，导出分数转换是为了将一个实测分数转换到一个可评价个体相对位置的分数系统上去。等值转换是两个或多个不同测验形式分数系统的转换，导出分数转换是一个测验形式不同分数系统的转换，两者之间是有本

质差异的。

寻找测验等值关系与寻找两测验之间预测关系也是不相同的。测验等值关系是测量同一种心理品质的多个不同测验形式、测验分数之间的转换关系，各个形式之间处于平等的地位。而预测关系两测验可以是测同种心理品质，也可以是测相近的甚至是不同的心理品质，预测源与预测目标之间的关系是不平等的，两者之间不是分数转换关系，它只能是从预测源的测试出发来预估预测目标的水平。

三、测验等值的条件

在两个不同形式的测验之间进行测验等值是必须具备一定条件的。测量学所提出的测验等值的条件主要有以下几个方面：

（1）同质性。被等值的不同测验形式所测的必须是同一种心理品质，测验的内容与范围也应该基本相同。不是测同一种心理品质的测验是不能被等值的。

（2）等信度。被等值的不同测验形式必须有相等的测验信度。不能指望一个低信度的测验通过与一个高信度测验等值而提高自身的可靠性。

（3）公平性。公平性是指：考生参加被等值的不同测验形式中的任一个的测试，等值后的结果都是一样的，不能出现参加不同形式的测试等值后的结果有高有低的现象。

（4）可递推性。如果测验 x 与测验 y 之间有等值转换关系 $f(x)=y$，测验 y 与测验 z 之间有等值转换关系 $g(y)=z$，那么一定有测验 x 与测验 z 之间的关系 h 存在，$h(x)=g(f$

(x)) = z。这种递推关系还可以推至更多的已等值的测验形式。如果这种递推关系不存在，或者不同途径递推的结果不相同，那么这些测验形式中必有不等值的形式存在。

(5) 对称性。对两个待等值的测验形式 x 与 y，无论等值转换从哪个测验出发，所获得的等值对应关系是相同的，即如果从形式 x 出发，获得等值关系 f (x) = y；从形式 y 出发，获得等值关系 g (y) = x，则一定有 $f = g^{-1}$，也就是说，f 与 g 一定是互逆的关系。

(6) 样本不变性。测验 x 与测验 y 的等值关系是由 x 与 y 的本身内在性质决定的，与为寻找这种等值关系而采集数据时所使用的样本没有关系，也与采集数据时测验的情境没有关系。如果测验等值关系会受到测试样本的影响而变化，则所寻获的测验等值关系是虚假的。

上述测验等值的 6 个条件，也有学者将前 4 条合称为公平性。在测验等值处理中，如果待等值测验能完全符合上述 6 个条件，则等值的结果将令人满意。但在实际研究中，可能会有个别的条件得不到满足，但并不完全否定等值的结果。比如，当测验形式要等信度的条件不满足时，在有些研究测验等值技术的专门文献中往往会给出另外一些补救的计算方法。但严格讲，这种方法已不能称为测验等值而被称为“测验校准”(Test Calibration)。

四、测验等值的一些基本概念

测验等值是一项综合性的测验统计分析技术，牵涉到测验理论的许多方面，也形成了许多专用基本概念，有些概念还常

常成对出现，为便于准确理解和应用，在此作一些介绍。

1. 经典理论等值与项目反应理论等值

两种等值的区别在于等值时以何种测验理论作指导。本章所介绍的等值方法均是在经典测验理论指导下的等值方法，这与本书的整体体系是一致的。但有研究者指出，应用经典理论等值，不满足等值条件的情况要更多一些，而应用项目反应理论等值，在等值条件方面会有较大的改善，从而使得等值的结果更为准确。

2. 测验分数等值与项目参数等值

这是根据测验等值的直接操作对象不同而构成的一对概念。如果等值的直接操作对象是测验的原始分数，结果是直接找到两测验分数的转换关系，称为测验分数等值。如果等值的直接操作对象是测验项目参数，找到的等值转换关系是两测验项目参数之间的转换关系，则称其为项目参数等值。项目参数等值可以是终极目的，但更多的是中间目的，在项目参数等值的基础上可以进一步找到测验分数的转换关系。为区别起见还是称其为项目参数等值。项目参数等值看上去似乎多了一道手续，实际上却很有用，特别是用于大型题库建设。利用项目参数等值可以把不同批次采集计算的项目参数，确定在一个统一的度量系统上，所有项目合并成一个大型题库。从这样的题库中抽题组成的不同试卷进行测试，可获得一致的评价结果，不必再进行等值计算。借助于项目参数等值而实现分数等值，其精度不比原始分数直接等值低。但必须指出的是，项目参数等值只有在项目反应理论的指导下才能进行。

3. 水平等值与垂直等值

这是根据测验试卷的难度和被试能力分布是否有差异而区分的一对概念。如果被等值的两测验形式有大体相同的难度水平，接受测验的两考生团体的能力分布也类似，这样两个测验

形式之间的等值称为水平等值。如果两测验形式的难度水平有明显差异，考生团体的能力水平也不相同，两个测验形式的等值称为垂直等值。显然，垂直等值的情况更为复杂一些，本书主要介绍的是水平等值的情况。

除了上述成对概念之外，测验等值中还有一些专用技术名词。

1. 测验等值设计

为了寻找不同测验形式之间的等值关系而预先对数据的采集方法、等值实现的途径、等值的计算方法进行周密的设计，称为测验等值设计。在实际工作中，并不是任何两个测验形式的原始数据都能用来进行等值计算的，两个测验形式分别施用于两个无关群体所获得的测验数据，就无法寻找到两形式之间的等值关系，因此在等值开始之时，就必须做好等值设计工作。在作等值设计时需要统筹考虑的问题包括：采用什么理论作指导、直接进行原始分数等值还是进行项目参数等值、等值数据如何采集、被试如何抽取、两测验形式之间以什么方法相关联、采集的数据用什么方法计算他们的等值关系等等。等值设计做得越科学，等值的效果就越好。

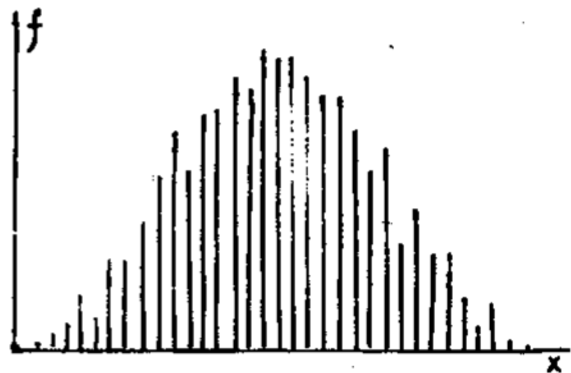
2. 锚（Anchor）测验

在测验等值设计中，有时会采用一组测验试题来关联两个待等值的测验形式，以便寻找两形式的等值关系，这些测验试题被称作为锚测验。锚测验在采集等值数据时，必须分别伴同两个待等值的测验形式向不同被试群体施测。锚测验可以嵌在原测验试卷中施测，也可以单独成卷与原测验分开施测。锚测验是嵌入试卷还是独立成卷要视数据采集的条件而决定，但不管施测形式如何，所起的作用是一样的。对于锚测验也是有一定要求的：锚测验也应与原测验一样测同种心理品质，锚测验也应与原测验有相同的测验信度，锚测验的长度一般不应小于

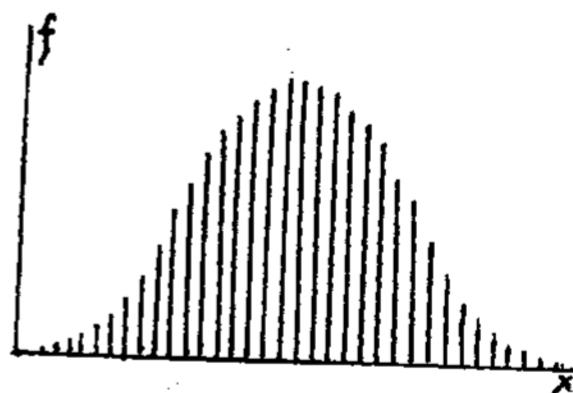
原测验的 1/5，理论上锚测验是越长越好，但不应造成被试的疲劳和厌倦。

3．数据平滑法

测验等值采集的数据来自于对样本群体的实测。由于条件限制，样本不可能很大，因此因此数据的稳定性不可能很理想，表现在分数分布中就是分布曲线的光滑性很差，特别是在分布的两端，由于被试的量少而使曲线波动较大（参见附图 9.1.A)。这对于要使用分数分布进行等值的计算影响很大，因此有必要对这种样本分布作一些技术处理，使得分布曲线趋向于比较光滑，统计上把这种技术称为数据平滑法。数据平滑的方法很多，测验等值中所用的数据平滑法中比较实用的有两种，一种叫对数线性平滑模式，一种叫 β 二项式平滑模式。这两种方法的应用均要涉及到一些较复杂的数学知识，我们在这里不作介绍，有兴趣的读者可以参阅相关文献。附图 9.1.B 是 9.1.A 经过平滑处理后的分布曲线。



A　平滑之前



B　平滑之后

附图 9.1　数据平滑示意图

4. 等值标准误差

测验等值的任何方法都要通过采集样本数据而完成计算、等值的结果肯定会受到抽样的影响而产生误差，测量学把由抽样而引起的等值误差称作等值标准误差。等值标准误差是可以用一定方法估计的，各种不同的等值方法有不同的等值标准误差的估计方法。测验等值标准误差是一个变量，随等值分数的大小而变，其总趋势是等值分数越趋于分布的两端，等值的标准误差就越大。

5. 等值偏差（Bias）

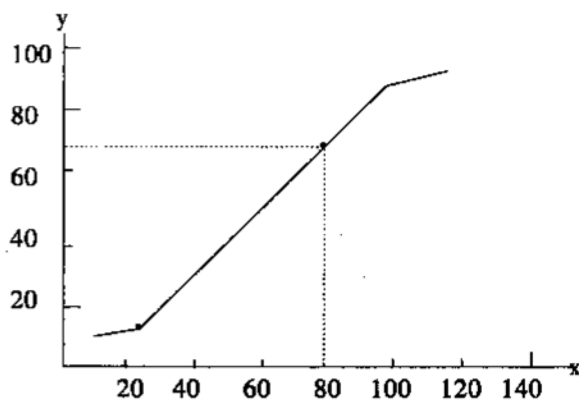
在测验等值中除了抽样引起等值误差之外，等值处理方法不当也会引起等值误差，测量学上把这种等值误差称为偏差。比如说，若是分别参加两测验形式测试的两个被试群体是两个能力有差异的独立群体，但被当成随机分组或等组处理，则等值结果就会产生偏差。在测验等值中，抽样引起的等值标准误差与处理方法不当引起的等值偏差常构成一对矛盾。比如在刚

才所举的例子中，为防止两被试群体能力不等而引起等值偏差，就必须增加锚测验测试，以估计能力差异大小，从而纠正等值偏差。但增加了锚测验又会增大抽样测试造成的误差，故需要研究者统筹考虑。

五、测验等值结果的表示方法

测验等值的结果是两个不同测验形式分数或项目参数间的转换关系，它的表示方法有3种。

第一种是表列法。将两形式对应相等的分数相对应排列成表，如附表9.2与9.3中所列。表列法简单明了，查找方便，是应用最普遍的等值结果表示方法。



附图9.2　x与y等值对应图

第二种是公式法。用于一些公式计算而获得的等值结果。

常见的等值结果公式形式为 $y = Ax + B$，其中 A 与 B 为等值常数，式中 x 与 y 是处于平等地位的。用公式表示等值结果简明、方便、等值关系清晰。但并不是所有的等值结果都能用公式表示的，而且公式法对于具体分数的配对还有一步计算要做。

第三种是图示法。如附图 9.2 所示，应用此图可以查找任一对等值的 x 与 y。

分数图示法形象生动地揭示了两测验分数间的等值转换关系，不受等值计算方法的限制。但图示法表示的对应关系精度有限，因此多用于对等值关系的整体分析。

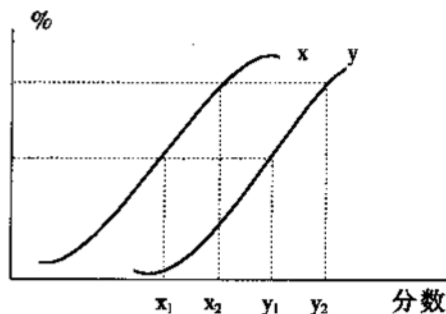
第二节　测验等值计算的基本方法

在经典测验理论指导下，测验等值的计算方法主要可以分为两大类：一类叫等百分位等值法，一类叫线性等值法。同一种等值数据采集模式既可用等百分位等值方法计算，也可以用线性等值方法计算。无论是等百分位等值计算方法还是线性等值计算方法随着数据采集模式的不同，具体的计算途径和公式也会有所不同。总的趋势是数据模式越复杂，计算途径和公式也就越复杂。但无论计算的具体途径和公式有什么不同，凡属于等百分位等值计算方法的或是属于线性等值计算方法的，其计算的基本原理都是一样的。为让读者对这两种等值计算方法的原理有所了解，本节就等百分位等值和线性等值两类计算方法各介绍些具体计算途径以示读者。

一、等百分位等值（Equipercentile Equating）

等百分位等值依据的原理是：两个分数，一个在测验形式 x 上，另一个在测验形式 y 上，如果这两个分数对于任何一个被试群体都有相同的百分等级，那么这两个分数就被认为是等值的。按照这个原理，寻找与 x 分数等值的 y 分数，只要找到与 x 分数有相等百分等级的 y 分数就可以了。

等百分位等值的关系寻找，可以通过将两测验名目的累积百分位曲线描绘在同一直角坐标系中获得（参见图 9.3）。图中的 x_1 与 y_1，x_2 与 y_2 就是成对的等值分数。这种方法称为作图法，显然作图法相对要粗糙一点。如果需要比较精确的结果，则可以在测验分数分布中应用百分等级计算公式求出与 x 分数对应的等值分数 y。我们用一例子来说明计算过程。



附图 9.3　等百分位等值

例 9.1　在某种等值设计之下采集得到两测验的分数、并

编制成次数分布表分列于附表 9.1 的 a 与 b，求这两个测验的等值分数对应表。

附表 9.1.a　x 测验分布

分组	f	F
90～94	7	302
85～89	19	295
80～84	27	276
75～79	33	249
70～74	42	216
65～69	45	174
60～64	39	129
55～59	31	90
50～54	24	59
45～49	18	35
40～44	13	17
35～39	4	4
合计	302	—

附表 9.1.b　y 测验分布

分组	f	F
85～89	8	290
80～84	11	282
75～79	18	271
70～74	24	253
65～69	32	229
60～64	40	197
55～59	45	157
50～54	39	112
45～49	28	73
40～44	23	45
35～39	16	22
30～34	6	6
合计	290	—

解：第一步：先分别求出两测验分数的向上累积次数分布，列于表末列。

第二步，设 $x=60$，在 a 表中求其百分等级。

$$PR=\frac{F_b+[(X-L_b)\cdot f/i]}{N}\times 100$$

$$=\frac{90+[(60-59.5)\times 39\div 5]}{302}\times 100$$

$$=31.0927$$

第三步：对已求 PR，在 b 表中求 y 分数。

$$y = L_b + \frac{\frac{PR}{100} \times N - F_b}{t} \cdot i$$

$$= 49.5 + \frac{\frac{31.0927}{100} \times 290 - 73}{39} \times 5$$

$$= 51.70$$

重复二、三两步，对所给出的任意 x 分数，都可求出与之等值的 y 分数，我们将部分等值对应分数列于附表 9.2 中。

附表 9.2　等百分位等值对应表

x	45	50	55	60	65	70	75	80	85
y	38.27	42.52	47.11	51.70	56.30	61.26	66.62	72.15	78.35
PR	6.2252	12.3841	20.5629	31.0927	44.2063	59.0066	72.6159	83.3444	92.0199

（解毕）

二、线性等值（Linear Rquating）

线性等值依据的原理是：两个分数，一个在测验形式 x 上，而另一个在测验形式 y 上，如果对于任何一个被试群体，它们各自的标准分数相等，这两个分数就被认为是等值的。线性等值原理如果用数学公式表示，所谓测验分数 x 等值于测验分数 y，即有下式成立：

$$\frac{x - \bar{x}}{S_x} = \frac{y - \bar{y}}{S_y} \qquad (9.1)$$

改写式9.1，成：

$$y = Ax + B \tag{9.2}$$

其中　$A = S_y/S_x$　　$B = \bar{y} - A\bar{x}$

如果能求出参数A与B，则对于测验x的任一分数均可利用式9.2求到与之等值的y分数。这里的A和B被称为等值常数。在线性等值中，两测验的等值关系为一直线，A是直线斜率，B是直线截距。所有的线性等值最终形式都是式9.2的形式，只是在不同的等值设计下A与B的求法不同罢了，此处列出的是最简单的计算A与B的方法。

我们用线性等值法来求例9.1提供的两测验分数分布的等值对应关系。

解：第一步：求出x测验分布的平均数与标准差

$$\bar{x} = 66.44 \qquad S_x = 12.98$$

第二步：求出y测验分布的平均数与标准差。

$$\bar{y} = 58.60 \qquad S_y = 13.05$$

第三步：求出等值常数A与B。

$$A = S_y/S_x = \frac{13.05}{12.98} = 1.0054$$

$$\begin{aligned} B &= \bar{y} - A \cdot \bar{x} \\ &= 58.60 - 1.0054 \times 66.44 \\ &= -8.1988 \end{aligned}$$

第四步：写出等值转换公式。

$$y = 1.0054x - 8.1988$$

对于给定的x，可用转换公式求出与之等值的y分数，我们将部分x的对应值求出列于附表9.3中。

附表 9.3　线性等值对应表

x	45	50	55	60	65	70	75	80	85
y	37.04	42.07	47.10	52.13	57.15	62.18	67.21	72.23	77.26

（解毕）

等百分位等值和线性等值是两种主要的等值计算方法，本节所介绍的都是两种计算方法中最简单的情况。在等百分位等值计算中，如果两个测验分数的分布形态相同，那么两测验形式的分数等值关系形成一条直线，此时的等值结果与线性等值的结果是一致的。但在大多数的情况下，两测验分数用等百分位等值求得的等值关系是非线性的。在实际应用中等百分位等值与线性等值的效果哪个更好要视具体的等值条件而论，很难作出绝对的结论。

第三节　常用测验等值设计介绍

• 设计一　随机分组——每组实施一个测验

对于两个待等值的测验形式 x 与 y，x 与 y 应测量同一种心理品质。选择一个充分异质的被减群体 t，将其随机分成两个被试组 α 和 β，$t=\alpha\cup\beta$，将测验 x 施测于被试组 α，将测验 y 施测于被试组 β。这样采集等值数据的方法称为随机分组——每组实施一个测验的设计。这一设计在两测验信度相等的条件下有两种等值计算方法。

1. 线性等值法

根据标准分数相等两原始分数等值的原理，可导出以下等值转换公式：

$$y = Ax + B \qquad (9.3)$$

$$A = S_{y\beta}/S_{x\alpha} \qquad B = M_{y\beta} - AM_{x\alpha}$$

其中 $M_{x\alpha}$、$S_{x\alpha}$为测验 x 施测于被试组 α 的平均数与标准差，$M_{y\beta}$、$S_{y\beta}$为测验 y 施测于被试组 β 的平均数与标准差，转换关系为一条直线，A 为直线斜率，B 为截距。

对于一个 x 通过 9.3 式转换而得的 y 值常记为 y^*，以便与实测值相区别。转换值 y^* 的等值标准误差记为 SE_{y^*}，标准误的平方可由下式近似计算：

$$SE_{y^*}^2 = S_{y_\beta}^2 \cdot \left(\frac{1}{N_\alpha} + \frac{1}{N_\beta}\right)\left(1 + \frac{1}{2}Z_{x\alpha}^2\right) \qquad (9.4)$$

其中：$Z_{x\alpha} = \dfrac{x - M_{x\alpha}}{S_{x\alpha}}$

2. 等百分位等值法

对于测验 x 的每一个分数，可据公式 9.5 在 x 测验分布中求出与其对应的百分等级 PR。

$$PR = \frac{F_b + \left[\dfrac{(x - L_b)\ f}{i}\right]}{N} \times 100 \qquad (9.5)$$

然后根据所求 PR 在 y 测验分布中用 9.6 式求出对应的 y：

$$y = L_b + \frac{\dfrac{PR}{100} \times N - F_b}{f} \cdot i \qquad (9.6)$$

当然也可以用前节所说的描图法求出各对等值分数，但两累积次数曲线应尽量修匀。

在本设计下等百分位等值的等值标准误差的近似计算公式为：

$$SE_{y^*} = [p \cdot q\ (\frac{1}{N_\alpha} + \frac{1}{N_\beta})]^{\frac{1}{2}} \Big/ f(y) \qquad (9.7)$$

其中 $p = P_R/100$，$q = 1 - p$，　$f(y)$ 为 β 群体中得分为 y 的人次数比。

• 设计二　随机分组——各测验对每组都实施

对于待等值的两个测验形式 x 和 y，所测的是同一种心理品质。假定形式 x 先测对 y 后测的影响与形式 y 先测对 x 后测的影响是相同的，选择一个尽量异质的被试团体 t，将其随机分成两个被试组 α 与 β，对被试组 α 先施测 x，再施测 y，对被试组 β 先施测 y，再施测 x，如此采集数据称为随机分组——各测验对每组都实施设计。如果两测验信度相等，也有两种方法可进行测验等值计算。

1. 线性等值法

根据标准分数相等的两个测验原始分数等值的原理，可导出以下线性等值公式：

$$\left.\begin{aligned} & y = Ax + B \\ & A = \sqrt{(S_{y\alpha}^2 + S_{y\beta}^2)\ /\ (S_{x\alpha}^2 + S_{x}^2\beta)} \\ & B = \frac{1}{2}\ (My\alpha + My\beta)\ - \frac{1}{2}A\ (Mx\alpha + Mx\beta) \end{aligned}\right\} \qquad (9.8)$$

此处所用各个符号的意义与 9.3 式中使用的完全一样。对于 x 测验的每一个分数，可用式 9.8 求出与之等值的 y 分数。在这一转换下，y^* 的等值标准误差的平方值为：

$$SE_{y^*}^2 = S_{yt}^2\ (1 - r_{xy})\ \frac{Z_{xt}^2\ (1 + r_{xy})\ + 2}{Nt} \qquad (9.9)$$

其中 r_{xy} 为两测验的积差相关，$Z_{xt} = (x - M_{xt})\ /S_{xt}$，$Nt = N_\alpha + N_\beta$。从 9.9 式可以看到，此设计的等值标准误受两测验的相关的影响较显著，从总体来看，此设计的误差比设计一的要小得多。

2. 等百分位等值法

本设计采集的实际资料是 x 与 y 两测验分别对整个被试群体 t 的测试资料。分组的目的仅仅是将两测验施测顺序的影响加以平衡。实施的结果是分别得到了 x 测验和 y 测验在全体被试施测的次数分布。注意这里的两个次数分布完全是同一批被试，因而有理由认为两百分等级相同的测验分数是等值的。计算过程与设计一的计算过程也是完全一样的。等值标准差的计算由于被试量的扩大而变小，感兴趣的读者可以自行推算出等值标准差的负增量。

• 设计三　随机分组——每组各实施一个测验，锚测验向每组实施

设计三与设计一相比，增加了一个锚测验向每个被试组实施，其目的是为了进一步控制两组被试的等价性，调整随机抽样后两组被试之间可能存在的差异。若记锚测验为 V，在 x 与 y 两测验信度相等的条件下，此设计可用线性等值法来完成等值计算，公式如下：

$$\left.\begin{aligned} y &= Ax + B \\ A &= \hat{S}_{yt}/\hat{S}_{xt} \\ B &= \hat{M}_{yt} - A \cdot \hat{M}_{xt} \end{aligned}\right\} \qquad (9.10)$$

上式中 $\hat{M}_{xt}$、$\hat{S}_{xt}$是 x 测验向全体被试施测时的平均数与标准差的估计值，$\hat{M}_{yt}$、$\hat{S}_{yt}$是 y 测验向全体被试施测时的平均数与标准差的估计值。因为 x 与 y 都未真正向全体被试施测过，因此这四项均只是估计值，估计公式如下：

$$\left.\begin{aligned} \hat{M}_{xt} &= M_{x\alpha} + b_{xv\alpha} \cdot (M_{vt} - M_{v\alpha}) \\ \hat{S}_{xt}^2 &= S_{x\alpha}^2 + b_{xv\alpha}^2 \cdot (S_{vt}^2 - S_{v\alpha}^2) \\ \hat{M}_{yt} &= M_{y\beta} + b_{yv\beta} \cdot (M_{vt} - M_{v\beta}) \\ \hat{S}_{yt}^2 &= S_{y\beta}^2 + b_{yv\beta}^2 \cdot (S_{vt}^2 - S_{v\beta}^2) \end{aligned}\right\} \qquad (9.11)$$

其中 $b_{xv\alpha}$ 为在群体 α 中测验 x 对测验 v 的回归系数，$b_{yv\beta}$ 为在群体 β 中测验 y 对测验 v 的回归系数。由式 9.10 估出的对 x 对应的等值分数 y^*，其等值标准误的平方值为：

$$SE_{y^*}^2 = 2\hat{S}_{yt}^2 (1-\hat{r}^2) \frac{Z_x^2 \cdot (1+\hat{r}^2) + 2}{N_t} \tag{9.12}$$

$\hat{r}$ 理论上可以是测验 x 与 v 或测验 y 与 v 的任意一个相关系数，实际应用中常取两者之均值。

• 设计四 非随机分组——每组各实施一个测验，锚测验向每组实施

在许多高度保密的测验中，采用两随机分组分别接受两个测验的设计会有许多实施上的困难。因而设计四将其修改为非随机分组，即允许两分组之间不是随机相等的。这样，即使两个分组的能力有所差异，采用本设计同样可以寻找到两测验的等值关系。下面的介绍还是在两测验信度相等的条件下进行。

1. 线性等值方法

在两被试组能力差异不大的情况下，线性等值的计算方法与设计三中所使用方法完全相同。若两被试组在能力上有差异，则 9.11 式中的 $\hat{M}_{xt}$，$\hat{S}_{xt}$，$\hat{M}_{yt}$，$\hat{S}_{yt}$ 改由下式估计：

$$\left.\begin{aligned}
\hat{M}_{xt} &= M_{x\alpha} + \frac{S_{x\alpha} \cdot \sqrt{r_{xx\alpha}}}{S_{v\alpha} \cdot \sqrt{r_{vv\alpha}}} \cdot (M_{vt} - M_{v\alpha}) \\
\hat{S}_{xt}^2 &= S_{x\alpha}^2 + \frac{S_{x\alpha}^2 \cdot r_{xx\alpha}}{S_{v\alpha}^2 \cdot r_{vv\alpha}} \cdot (S_{vt}^2 - S_{v\alpha}^2) \\
\hat{M}_{yt} &= M_{x\beta} + \frac{S_{y\beta} \cdot \sqrt{r_{yy\beta}}}{S_{v\beta} \cdot \sqrt{r_{vv\beta}}} (M_{vt} - M_{v\beta}) \\
\hat{S}_{yt}^2 &= S_{y\alpha}^2 + \frac{S_{y\alpha}^2 \cdot r_{yy\beta}}{S_{v\beta}^2 \cdot r_{vv\beta}} (S_{vt}^2 - S_{v\beta}^2)
\end{aligned}\right\} \tag{9.13}$$

其中 $r_{xx\alpha}$，$r_{vv\alpha}$，$r_{yy\beta}$ 与 $r_{vv\beta}$ 均为测验的信度。估出上述四值之

后，代入 9.10 式可求到两测验的等值关系式。

2. 频数估计法

频数估计法（Frequency Estimation Method）是等百分位等值的一种，用于有锚测验的等值设计。频数估计法的关键是要利用锚测验数据分别估出测验 x 和测验 y 在合成被试群体 t 上的次数分布。获得了两测验的次数分布就可用设计一所提供的等百分位等值方法求出测验 x 与测验 y 的等值对应关系了。

估计 x 测验和 y 测验在合成总体 t 上的次数分布的方法是一样的，我们以估计合成总体在 x 测验上的次数分布为例来演示这一方法。

设 α 被试组参加了测验 x 和锚测验 V 的施测，其在两测验上的联合次数分布如附图 9.4 表一所列。β 被试组只参加了 V 测验施测而没有参加 x 的施测。因此只有在 V 测验上的次数分布是已知的（参见附图 9.4 表二　中的最右列），我们的

9.4 表一　α 被试组在 x 与 v 上的联合分布 f

v \ x	0	1	2	3	4	5	合计
0	0	1	0	0	0	0	1
1	0	0	1	1	0	0	2
2	0	0	1	2	1	0	4
3	0	0	0	2	4	2	8
合计	0	1	2	5	5	2	15

9.4 表二　β 被试组在 x 与 v 上的联合分布（估计）g

v \ x	0	1	2	3	4	5	合计
0	0	2	0	0	0	0	2
1	0	0	2	2	0	0	4
2	0	0	1.5	3	1.5	0	6
3	0	0	0	1	2	1	4
合计	0	2	3.5	6	3.5	1	16

附图 9.4　频数估计示例

第一项任务是完成表二中联合分布的估计。估计的思路是：若α组在V测验上有$f_{i.}$人得i分，β组在V测验上有$g_{i.}$人得i分，由于这两部分人在V测验上得分相等，因此认为这两部分人在x测验上也应有相同的得分分布。即：若α组中在V测验上得i分的$f_{i.}$人中有f_{ij}人在x测验上得了j分，那么就认为β组中在V测验上得i分的$g_{i.}$人中应有$g_{ij}=g_{i.}\cdot f_{ij}/f_{i.}$人在x测验上得j分。比如，β组在V测验上得2分，在x测验上得4分的被试人数估计值$g_{24}=g_{2.}\cdot f_{24}/f_{2.}=6\times1\div4=1.5$。根据这一思路，我们可估出表二中的所有$g_{ij}$，进而纵向累计成β组被试在x测验上的次数分布（附图9.4表二末行）。第二项任务就是把α组在x测验上的分布（实测）与β组在x测验上的分布（估计）合成为总体t在x测验上的分布。其分布列在附表9.4中。

附表9.4　合成总体在X测验上的分布

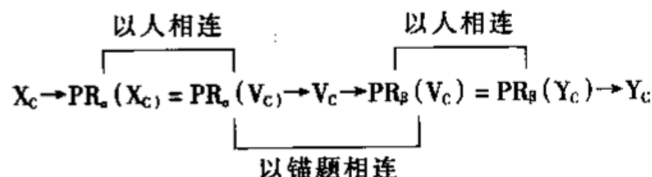
x	0	1	2	3	4	5	合计
f	0	3	5.5	11	8.5	3	31

按照同样的办法，我们可以估出合成总体在y测验上的得分分布。这样我们就可以用设计一提供的等百分位等值法将x与y两测验等值了。

3. 链等值法（Chained Equipercentile Equating Method）

链等值法采取的是一种链接传递的等值计算策略。在本设计中，先利用α被试组既参加了x测验又参加了v测验的条件，应用等百分位等值法将测验x分数与测验v分数等值对应，再利用β被试组既参加y测验又参加了v测验的条件，应用等百分位等值将测验y分数与测验v的分数等值对应，那么

通过两次等值，又通过锚测验 v 的链接，测验 x 与测验 y 也就有了等值对应关系。在操作上可以遵循以下途径：对于测验 x 的某个分数 x_c，在 α 被试组的 x 测验次数分布中求出对应的百分等级 PR_α（x_c），其值应该与 α 被试组在 v 测验上某个 v_c 分数的百分等级 PR_α（v_c）相等，据已知的 PR_α（v_c）在次数分布表中可求得与 x_c 等值的 v 测验分数 v_c，据 v_c 分数，根据 β 被试群体在 v 测验上的分布，可求出相应的百分等级 PR_β（v_c），其值应该与 β 被试组在 y 测验上某个 y_c 分数的百分等级 PR_β（y_c）相等，据已知的 PR_β（y_c），由 β 组在 y 测验上的次数分布可求得与 v_c 等值的 y 测验分数 y_c，故而有 x_c 与 y_c 对应等值。整个过程示意如附图 9.5。



附图 9.5　链等值示意图

通过图 9.5 可以看到，测验等值关系的寻找一定要通过某种等值媒体。这个等值媒体，或者是同一批被试（或认为能力分布相等的两批被试），或者是同一批测题（锚题或锚测验）。在等值设计中这是一个必须保证的条件。另外还必须指出的是，应用等百分位等值进行等值计算最好是先对样本次数的分布作平滑处理，以获取最佳等值效果。

练习与思考

1. 应用例9.1的数据采取等百分位等值方法完成下面的等值对应表。

x	35			90
y		50	70	

2. 若例9.1数据是设计一之下采集获得的，请分别求取与 x = 65 等值的 y 测验分数的两种等值标准误（等百分位等值与线性等值两种）。

3. * 求取两平行测验之间的回归方程，可以建立两测验分数之间的对应关系。这种对应关系能不能称为等值关系，为什么？

4. * 实际采集一批数据对两个测验进行等值计算。

第十章 目标参照测验

本章提要：

- 目标参照测验的特殊意义
- 目标参照测验的项目分析
- 目标参照测验的信度与效度分析
- 目标参照测验合格分数分界点的确定

第一节　概述

一、目标参照测验的产生

20世纪初期，科学的测量方法引入心理学和教育学的研究领域。出于解释测验原始分数含义的需要，西方早期的心理与教育测量学家们提出“相对能力”的测量，即将测验的原始分数转换为百分等级或标准分数等相对位置量数，从而指出个体在其相应团体中相对于其他个体而言的能力水平。此后，心理与教育测量的一个基本出发点就是度量与比较个体间的差异，以被试在团体中的相对位置来评定和解释测量结果。在这一基础上逐渐发展起来的一个比较固定的测验模式就是常模参照测验，它广泛应用于能力、能力倾向、成就、人格、态度等多种特质的测量之中，并发展起一套比较成熟的统计分析方法，用于项目分析、质量控制（测验信、效度分析）以及分数解释之中。

然而，从本世纪中期开始，人们逐渐发现常模参照测验模式存在一定的局限性：并不是所有的测验都只关心个体间的差异。有些测验目的在于了解和界定个体在测验内容上掌握的绝对水平。比较典型的例子是用于评价教学活动结果的测验，它们的目的是为确定在某一特定教学领域内，被试是否掌握了该领域中必要的知识或技能以及他在这一领域中的困难与缺陷所在，以便有目的地对他加以教学辅导与补救。常模参照测验则

只能描述被试在团体中的相对位置，无法说明他对所测内容掌握的绝对水平，因而这种测验模式在此便显得很适用了。

鉴于常模参照测验的这种局限性，测量学家们开始考虑另一种可供选择的模式：目标参照测验（亦称标准参照测验）。1962年戈莱塞（R.Glaser）和克劳斯（D.Klaus）首先提出目标参照测验的概念，并于次年详细论述了这一测验模式在成就测验上的功用，从而使得目标参照测验引起测量学家的极大关注，并逐渐发展成与常模参照测验并列的一种测验模式。

二、目标参照测验的定义

目标参照测验的主要目的在于了解个体在所规定的测量内容上的行为水平，因此其出发点是个体本身的绝对水平，而不再是个体间的水平差异。对于这样一种不同于常模参照测验的新的测验模式，测量学者们从不同角度给它下了许多不同的定义，至今还没有一个统一的为众人公认的结论。但就一般意义而言，戈莱塞1971年对于目标参照测验的描述与界定是比较广泛地为人们所接受的："所谓目标参照测验，是根据某一明确界定的内容范围而缜密编制的测验，并且，被试在测验上所得结果，也是根据某一明确界定的行为标准直接进行解释的。"

在这一定义中，"内容范围"的概念是首要的，在测验编制之前，必须对所欲测量的内容范围做出清晰的界定，并给予它严格的操作定义。测验题目的选择限制在这样的内容范围之内，并且，构成测验的所有题目，必须是所依据的内容范围的一个代表性样本。这样一来，被试在测验中的成绩，便有理由被推论到测验所欲测的内容范围中去，从而可以对被试在所测

内容范围中的掌握水平作出评价。

"行为标准"是上述定义中的另一个重要概念。目标参照测验的目的一般在于了解被试在某一行为领域的绝对水平，从而判定他是否达到了从事此项行为的最低标准，比如中学会考的目的在于判断考生是否达到了中学毕业所要求的最基本的知识技能水平，各种专业化的资格考试目的在于考察考生是否具备了从事这一专业所要求的最低水平，等等。因此，目标参照测验的分数一般将依据某一绝对的标准进行解释，这一标准一般称为"分界点"。并且，目标参照测验的分界点的确定，是建立在内容范围的明确界定基础之上的（详见本章第四节）。

因此，在一般的意义上，当一个测验是以某一明确界定的内容范围为基础编制而成，并且其分数是参照该内容范围所要求的绝对标准进行解释，我们便称这一测验为一个目标参照测验。

第二节 目标参照测验的项目分析

一、内容范围的确定

任何一种测验的编制，其前期工作不外乎为测验目的的确定，测验内容的界定以及测验编制计划的设计。从目标参照测验的定义来看，构成测验的各个项目是否合适，测验是否有效，测验的分数是否能得到有意义而准确的解释，这一切的前提都在于测验有没有明确的目的以及与之相应的严格界定的内容范围。因此，对于目标参照测验而言，测验编制的前期过程

尤为重要。

一个测验的内容范围包括所欲测量特质中蕴含的全部行为，它可以非常大，如数学能力，也可以非常小，如10以内的整数加法运算能力。不过，任何一个内容范围都具有一些共同的特点。首先，内容范围具有边界。当其边界得到明确界定时，我们就可以判定什么行为属于这一内容范围，而什么行为却超出了这一范围。其次，每一内容范围内容均可分为几类，每一类中又可分为更细更小的类，当每一类的内容及其在此内容范围内的相对重要性确定以后，内容范围就有了明确的结构。而当一个内容范围具有了明确的边界和结构时，我们便认为此内容范围得到了明确界定。

特定测验目的的确定常为内容范围的界定提供依据。如若测验目的在于检验某类专业化工作的资格水平，那么通过工作分析便可界定测验的内容范围；若测验的目的在于检验教学或训练的效果，那么可以通过与特定课程或训练有关的教材、大纲以及学科专家的意见来界定内容范围。界定的结果常常以双向细目表（或称测验蓝图）形式表现出来。

表10.1是广东省化学高考标准化试验中使用的命题细目表（1986~1990年）。

如上例所示，命题细目表由3个要素构成：一是教学目标，本例中列有识记、理解、应用、分析综合与评价五个方面；二是教学内容，一般可参照本学科的教学大纲和教材来确定；三是在整个内容范围中每一类内容和每一种目标相结合后所占的比重（相对重要性），上表中数字即为比重值。这一要素主要通过专家评定而获得。

表 10.1 命题细目表举例

考查内容 \ 考查目标	识记	理解	应用	分析综合	评价	合计
基本概念与理论	1	13	9	7	2	32
元素化合物	3	5	6	5	2	21
有机化合物	1	5	3	4	2	15
化学计算	0	3	4	8	0	15
化学实验	1	6	2	6	2	17
合计	6	32	24	30	8	100

在上例所示的命题细目表中，化学高考的内容范围已具备了明确的边界和结构，试卷的编制工作便可在这一框架中进行。

二、测验项目的内容效度分析

目标参照测验的项目分析，首先要对构成测验的每一个题目是否合适以及有效进行分析，即检验题目与测验内容范围所要求的内容与目标的一致性。这一过程一般缺乏客观的统计分析手段，通常采用专家评定的方法。

专家评定可以采取不同方式，其中比较直观和常用的一种方式是要求有关内容领域的专家填写项目内容评定表，在五级量表上对每个题目所测内容与项目编制者所欲测量的目标内容之间的一致性作出评定，表 10.2 是一个测验项目内容评定表的样例。

表 10.2　项目内容评定表（样例）

项目内容评定表

评定者姓名：　　日期：　　内容范围：

首先，请仔细阅读已界定的内容范围和测验项目；

然后，请判断：你认为每一项目在多大程度上反映了其在被编制时所欲测的目标内容。判断赖以产生的唯一基础是项目内容与其意欲测量的目标内容之间的匹配程度。请采用下面的五级量表：

较差匹配	一般匹配	较好匹配	很好匹配	完美匹配
1	2	3	4	5

在测验项目的题号所对应的项目评定栏中你认为合适的等级数目上划圈。

目标内容	测验题号	项目评定
1	2	1　2　3　4　5
	7	1　2　3　4　5
	14	1　2　3　4　5
2	1	1　2　3　4　5
	3	1　2　3　4　5
	8	1　2　3　4　5
	13	1　2　3　4　5
3	4	1　2　3　4　5
	6	1　2　3　4　5
	12	1　2　3　4　5
4	5	1　2　3　4　5
	9	1　2　3　4　5
	10	1　2　3　4　5
	11	1　2　3　4　5

注：此表引自 Ranald A.Berk 所著《目标参照测验导论》。

对于测验中每一项目的内容与其目标内容之间一致性的等级评定，通常需要邀请多位专家共同进行，这样便可以得到多位专家的评定结果，表 10.3 是 9 位专家在表 10.2 所示评定表中的等级评定结果以及对此结果的一些统计数据。

表 10.3　9 位专家对 14 道题目等级评定结果

目标内容	测验题号	专家评定结果 1	2	3	4	5	6	7	8	9	统计数据 平均数	中数
1	2	4	3	5	5	4	5	5	5	4	4.4	5
	7	4	2	5	5	5	5	5	4	5	4.4	5
	14	4	5	5	5	4	5	5	5	5	4.8	5
2	1	3	5	3	2	1	4	5	2	4	3.2	3
	3	3	1	4	4	3	4	4	3	3	3.2	3
	8	1	3	1	2	1	1	1	1	1	1.3	1
	13	1	3	2	1	1	2	1	2	3	1.8	2
3	4	4	5	5	4	5	5	5	5	5	4.8	5
	6	4	2	4	4	4	4	4	4	4	3.8	4
	12	5	3	5	5	5	5	5	5	5	4.8	5
4	5	4	3	5	5	4	5	5	4	5	4.4	5
	9	2	2	4	1	4	2	4	4	4	3.0	4
	10	1	3	1	2	1	1	1	1	1	1.3	1
	11	4	3	4	4	5	5	5	5	5	4.6	5
专家判断与中数的差异		9	24	2	10	6	4	4	3	3		

注：表中资料来源同表 10.2。

表 10.2 不仅列出 9 位专家对 14 道测验题目分别作出的等

级评定结果，而且还给出了九位专家在每道题目上的等级评定均值和中位数，以及每位专家对 14 道题目所评等级与中数的差异之和。

从表 10.3 的结果中，我们可以直接分析每一道题目的内容效度并进而决定题目的取舍。若以中数为基础进行分析，表中第 2、7、14 题均是针对目标内容 1 而编制的，经专家评定，这三题所测内容均和其目标内容之间具有完美的匹配（等级评定中数均为 5）；同理，第 4、12、5、11 题的内容效度亦得到很高的评价。因此，这七道题目无疑可以原封不动地保留下来。再看，第 1、3、6、9 题所测内容与其目标内容之间的匹配程度分别被评定为较好（中数为 3）或很好（中数为 4），说明这四题也还是可取的，只需根据专家意见略作修改即可，最后，第 8、13、10 题的内容效度一般（中数为 2）或较差（中数为 1），说明这三题没有能够很好地反映出其欲测的目标内容，内容效度很低或根本不具备内容效度，一般需作较大的修改或删除。

若以均值为基础进行分析，得到的结果与上述以中数为基础进行分析的结果是十分类同的。有时，为了增加参加评定的专家们之间的一致性，也可根据每位专家在所有题目上所评等级与中数间的差异量来决定专家的取舍。如上表所示，第二位专家在所有题目上所评等级与各中数间的差异之和为 24，说明该专家的评定结果与其他 8 位专家之间具有较大的差异，因此可以不考虑这位专家的意见，只保留 8 位专家的评定结果，在此基础上得到均值和中数等统计数据并进一步决定题目的取舍。

三、测验项目的难度和区分度分析

（一）测验的预测

测验编制完成后，须选取一定数量的被试进行预测，由此获得预测数据，然后才能在此数据基础上对项目的难度和区分度进行量化分析。

目标参照测验的预测方法主要有以下三种：

1. 前测—后测方法

选取一组被试，在其接受与测验目标内容有关的教学过程前后各施测一次，取得前测和后测的结果，前者表示未掌握者在测验中的水平，后者表示已掌握者的水平。

2. 已接受教学组——未接受教学组方法

选取两组被试，其中一组已经接受了有关测验目标内容的教学，而另一组从未接受过，将测验对这两组被试同时施测，亦可获得与第一种方法中含义类同的两组结果。

3. 对照组方法

方法 1 和 2 均假设凡接受了有关教学活动的被试均已掌握了教学内容，因而视之为掌握组。然而，在实际当中，很可能在已接受有关教学的被试中依然存在个别未掌握者，而在从未接受有关教学的被试中却存在个别掌握者，因而方法 1 和 2 在这一点上是值得质疑的。对照组方法的提出可以说是对此缺陷的弥补：选取两组被试，其中一组被试被其教师评定为掌握组，而另一组则被教师评定为未掌握组。将测验同时施测于这两组被试，便获得与上述方法类同的结果。

（二）测验项目的难度分析

目标参照测验的项目难度计算与常模参照测验相同，一般以通过率来表示。但是，有些学者认为，目标参照测验的项目难度分析并不重要，甚至有时并不必要。纯粹的目标参照测验一般注重的是所测内容范围以及被试在所测内容范围上的掌握程度，因而若某项目所测为内容范围内不可或缺的重要内容，那么无论该项目是难是易，均应得到保留。

对于目标参照测验的项目难度的计算，在大多情况下只是作为项目区分度分析的基础。

（三）测验项目的区分度分析

目标参照测验应该能将在其内容范围上的已掌握和未掌握者作出最大限度的区分，因而，每一测验项目的区分度如何便成为一个值得关心的问题。

测验项目的区分度一般采取两类指标：难度差值和相关系数。

1. 难度差值

（1）掌握组——未掌握组鉴别指数（D）

通过上述三种预测方法中的任何一种，均可得到两组数据，一组代表掌握者水平，一组代表未掌握者水平。分别计算这两组在某项目上的平均通过率，记为 P_A 和 P_B，则该项目的鉴别指数为：

$$K = P_A - P_B$$

鉴别指数 D 的大小，可以直观反映出该项目在多大程度上对掌握者和未掌握者作出了区分。D 值从 - 1.00 到 + 1.00 之间变化，越接近于 + 1.00，题目区分度越高，说明题目越有效。以表 10.4 中数据为例，可对表中五个项目的区分度进行分析。

表 10.4　前后测的项目得分表

被试	项目									
	1		2		3		4		5	
	前测	后测	前测	后测	前测	后测	前测	后测	前测	后测
1	0	1	1	0	1	1	1	1	1	1
2	0	1	1	0	1	1	0	1	1	1
3	0	1	1	0	0	1	1	1	1	1
4	0	1	1	0	0	0	1	1	1	1
5	0	1	1	0	1	0	0	1	1	1
6	0	1	1	0	0	0	0	0	1	1
7	0	1	1	0	1	0	1	1	1	1
8	0	1	1	0	1	1	0	0	1	1
9	0	1	1	0	0	1	0	1	1	1
10	0	1	1	0	1	0	0	1	1	1

表中后测分数表示掌握者分数，前测分数则表示未掌握者分数，则：

对于项目 1，$D_1 = 1 - 0 = 1$

同理，$D_2 = 0 - 1 = -1$

$$D_3 = \frac{5}{10} - \frac{6}{10} = -0.1$$

$$D_4 = \frac{8}{10} - \frac{4}{10} = 0.4$$

$$D_5 = 1 - 1 = 0$$

分析这五个项目的区分度值，D_1 为 +1.00，说明项目 1 可以将掌握者和未掌握者作出最准确的区分；$D_2 = -1.00$，说明项目 2 虽然也将掌握者和未掌握者作了最大区分，但问题在于掌握者无人通过该题，而未掌握者却全部通过，可见该项目存在错误，或者是出题错误，或者是正确答案弄错，此时应

仔细查找该题错误原因，予以修改或删除；项目 3 的鉴别指数亦为负值，即未掌握者在该题的通过率高于掌握者，说明出现与项目 2 类同的问题，因此应对此项目作出类似对项目 2 的处理；$D_4=0.4$，说明项目 4 已具有一定的区分度，可以保留；$D_5=0$，说明项目 5 对掌握者和未掌握者具有同样的难度，亦即不具备区分力，在纯粹目标参照测验中，若该项目所测内容非常重要，那么仍可考虑保留该题。

(2) 个人获得指数（D_{IG}）

采取前测—后测方法，可获得在前测中错误回答某项目而在后测中能够正确回答的被试人数比例，此即该项目的个人获得指数，其值在 0 至 +1.00 之间变化，其大小直接反映了经过教学活动之后受益的被试比例。以表 10.4 中数据为例，可以得到五个项目的个人获得指数分别为 1，0，0.2，0.4，0，说明项目 1 最有效，项目 2 和 5 没有区分力。

由于 D_{IG} 只考虑到前测中失败而在后测中通过的被试，却没有考虑到在前测中通过而在后测中反而失败的被试，因而其值不会出现负值，这使它所能反映的问题少于一般的区分度指标，因此其应用也受到限制。

2. 相关系数

项目得分和测验总分之间的一致性程度常被用作项目区分度的指标，这在常模参照测验的区分度计算中已有详细介绍，这些方法在目标参照测验中同样适用。

以相关系数为指标的区分度在 -1.00 至 +1.00 之间变化，当其值为负时，应检查题目的问题所在，予以修改或删除；当其值为正时，越接近于 1，题目越有效；当其值为 0 时，题目不具区分力，一般不予保留，但在纯粹目标参照测验中应视该题所测内容在内容范围中的重要性而决定其取舍。

目标参照测验的项目区分度还可以用其他许多指标进行计

算和分析，但由于这些指标所需计算比较复杂，因而在实际中并不常用，在此不再介绍。

第三节 目标参照测验的信度与效度

一、信度及其估计

信度是指测量结果的一致性或稳定性。任何类型的测验，都应该保证测验结果的信度，即对同一施测对象施测多次后的结果之间应该具备高度的一致性，从而可以将测验结果归之于个体真实水平的影响而非随机误差的影响，对目标参照测验的质量评估同样应该重视信度这一指标。

在常模参照测验的信度评估中，通常是以相关系数作为信度指标，相关程度越高，信度就越高，测验越可靠。然而，由于以相关系数表示的测验信度的高低在很大程度上受到受测者团体异质性的影响，即被试异质性越高，测验分数分布就越广，从而相关系数越高，测验信度相应也就越高，因而，这些信度指标在目标参照测验上的应用价值就不免受到怀疑：目标参照测验的目的一般不在于鉴别个体差异，而在于了解个体在所测内容上的掌握水平，因而在大多数情况下，被试团体在目标参照测验上的分数分布比较集中，如高中毕业会考，一般来说绝大多数考生都能达到所要求的水平。这样一来，若用相关系数作信度指标，由于其受到分数分布的影响，那么即使测验本身具有较高的稳定性和一致性，所得的信度系数也会很低。

可见，通常以相关系数所表示的信度指标在目标参照测验上是不太适用的。

对于目标参照测验的信度估计，测量学家们正在不断地探索着适宜的统计方法，有些人也提出了一些统计指标，不过还不够成熟。现介绍两种方法如下。

(一) 分类一致性信度

目标参照测验在其分数解释上最常见的做法就是将被试分类，一般是根据某一分数分界点将被试分为掌握者和未掌握者两类。称作“达标—未达标”或者“及格—未及格”。因此，分类的一致性在此就显得非常重要。

对测验的分类一致性的度量指标，称作为分类一致性信度。其最简单易行也最常用的估计方法是考查被试在同一测验的两次施测中或两个复本的施测中是否被分在同一类中。具体做法类似于常模参照测验中的再测法和复本法，但统计方法和所用指标不同。在此是采用同一被试团体在两次测验结果中均被分为及格或不及格类别中的人数百分比例作为分类一致性信度的指标。

设两次测验记为 A 和 B，测验结果以表 10.5 表示：

表 10.5　两次测验结果分类表

		测验 A	
		及格	不及格
测验B	及格	a	b
	不及格	c	d

据上表，测验的分类一致性信度为两次施测中均及格和均不及格人数占总人数的比例，即：

$p_0 = (a+d)/N$

式中，$N = a+b+c+d$

若有一个60人的团体，在某测验的两次施测中有21人均及格，12人均不及格，那么，该测验的分类一致性信度则为 $p_0 = (21+12)/60 = 0.55$。

分类一致性信度 p_0 的最大值为1，说明两次施测结果对被试的分类完全一致，测验结果完全一致。p_0 越接近于1，说明测验结果的一致性或稳定性越高，测验越可靠。

分类一致性信度的优点在于计算简单，意义直观易懂。但由于其所采用的方法类同于常模参照测验信度评估中的再测法和复本法，因而再测法和复本法的缺点在此同样存在。此外，分类一致性信度也受到测验长度和被试分数分布的影响。不过，在分类一致性信度的影响因素中，最重要也最独特的一个因素是测验分数分界点的确定问题。分界点不同，意味着被划分为及格和不及格的标准改变，因而人数比例必然也会发生变化。因此，分数分界点的科学确定是评估测验分类一致性的前提（分界点的确定详见本章第四节）。在报告目标参照测验的分类一致性信度时，必须同时提供测验的分数分界点，以及测验长度等资料。

在目标参照测验的分类一致性研究中，还有人提出其他一些指标，但它们或者由于计算过分复杂，或者由于解释不够直观，在应用上一直不如 p_0 广泛。

（二）方差分析方法——荷伊特信度

在经典测量理论的真分数模型中，信度被定义为真分数的

变异在实得分数变异中所占比例。常模参照测验的信度评估方法中所介绍过的荷伊特信度，正是从信度定义出发，利用方差分析的方法，找出个体水平的真正变异在总变异中的比例，以此作为信度的估计值。此法不受测验目的或被试异质性的影响，因而同样适用于目标参照测验的信度评估。此法具体计算与解释在本书前文已有叙述，在此不再赘述。由于荷伊特信度不会随测验分数分界点而变化，因而更具普遍性。

二、效度及其估计

测验的效度是评价任何一个测验的质量好坏的最重要的指标，因而，效度分析在目标参照测验的质量评估中同样占据重要地位。

（一）内容效度

目标参照测验注重被试在其所测内容范围内的掌握程度，因而测验本身的题目组成对其欲测之内容范围的覆盖程度或代表性程度——亦即测验的内容效度——在此显得尤为重要。

评估任一测验的内容效度，都依赖于两个条件：一是测验有明确界定的内容范围；二是对测验每一题目的内容效度的分析。目标参照测验一般来说有相对比较确定的内容范围，可以命题细目表表示，同时，也可以采用专家评定的方法对题目效度进行分析，从而保留有效题目，删除无效题目。下一步的问题便在于：所有保留下来的有效题目对整个内容范围的覆盖程度如何？对此，常模参照测验中所介绍的内容效度分析方法基本上可以照搬到目标参照测验中来，在此不再赘述。

（二）效标关联效度

目标参照测验一般倾向于根据被试在测验中所得分数将其划分至掌握者或者未掌握者之中，从中可以对被试在未来的学习或工作上可能成功的程度作出预测，从而为教学决策或一些人事方面的决策提供依据。因此，效标关联效度（也称实证效度）的分析对目标参照测验来说也是重要的。

目标参照测验的效标关联效度分析方法与常模参照测验中所介绍的方法在具体实施中没有太大差异，其不同之处主要在于统计指标上。常模参照测验一般用测验与效标间的相关系数作为测验效标关联效度的指标，而我们已经知道，相关系数大小受到分数分布的影响，不适用于目标参照测验，因而有人提出以"决策效度"（Decision Validity）来评估目标参照测验的效标关联效度。

以教学情境中某目标参照测验为例：测验结果依据某分数分界点分为及格和不及格两类；选用"是否接受过相应教学活动"或者教师评定结果为效标，将参加测验的被试分为"掌握组"和"未掌握组"；计算掌握组被试在测验中及格人数占参加测验总人数的比例和未掌握组在测验中不及格的人数比例；两个比例相加所得结果即为决策效度。

决策效度的计算方法也可以类似表 10.5 的形式来表示，只不过在分类一致性信度的计算中，表 10.5 中测验 A 和 B 是指同一测验的两次施测或等值的两个复本，而在这里的效度计算中测验 A 和 B 一是指预测源测验，另一是指效标测验，而所谓决策效度即指在预测源测验和效标测验中均通过和均不通过的被试人数百分比例。

由于目标参照测验在多数情况下是对于被试在特定教学或训练内容上的掌握情况的检查，因而人们较少关注测验目的所

蕴含的理论构想问题。况且，常模参照测验的结构效度评估大多是以相关系数为基础，不适用于目标参照测验。所以，关于目标参照测验的结构效度，目前尚未得到较大关注。

第四节　测验分数的解释
——分数分界点的确定

回顾目标参照测验的定义，其测验结果是参照某一明确界定的行为标准进行解释的，这一标准就是测验分数的分界点，亦称切割分数线，或称及格线。根据分数分界点，可以将被试进行分类，通常分为“及格”和“不及格”两类。在这样的分类过程中，分界点的确定是至关重要的。

事实上，就目标参照测验本身而言，分数分界点并非必需。我们可以用“被试掌握了测验的内容范围的百分之多少”来解释被试的分数，而不必在测验分数这一连续体上寻找某个切割点，进而将被试断然分为两类：掌握者或非掌握者。一般来说，人们倾向于认为知识的学习是一个连续的过程，知识的掌握也只是一个程度的问题，因而从理论上说并不存在可以清晰辨别的掌握者或“非掌握者”。这使得分数分界点的确定成为测量学家们争议最大，存疑最多的问题。

然而，在目标参照测验的实际应用中，分数分界点的确定却是无法逃避的问题。在教育领域，我们常常需要根据测验结果来判断：“某学生是否达到了升一个年级（或小学、初中、高中、大学毕业等）所要求掌握的最低知识技能水平”，从而对该学生“升级”或“留级”、“毕业”或“肄业”等作出决

策；在专业领域，也常需要根据资格或水平考试结果来判断考生是否达到从事特定专业工作所需的最低水平，从而作出是否给予颁发合格证书的决策。在这些实际需要中，我们不得不去寻找一个最低标准，一个分数分界点或及格线，将考生分为及格或不及格两类。而且，这一分界点的确定科学与否，直接决定了我们最终决策的正确与否。因此，探索分数分界点的确定方法是必要而且重要的。

迄今为止，测量学家已经提出了许许多多的分数分界点的确定方法，这些方法各有利弊。现介绍其中比较常用的几种方法如下。

一、专家判定法

这种方法是在测验的内容范围明确界定的基础之上，由专家来判断处于临界水平的被试在每一题目上正确回答的可能性，进一步以此为标准确定分数分界点。所谓临界水平的被试，是指那些刚由未掌握水平转入掌握水平的被试，这些被试实际上是在专家的想象中虚拟出来的。

具体评定方法主要有以下两种：

（一）Nedelsky 方法

此法由 Nedelsky（1954）提出，针对由多重选择题组成的测验而言，由专家来判断处于临界水平的被试在每一题上有能力排除的错误选择项，从而计算其正确回答的可能性，再求出每一题上正确回答的可能性之和，即为测验分数分界点。例如，某测验由四择一选择题组成，某题 A、B、C、D 四个答

案中 A 是唯一正确的答案。若专家判定处于临界水平的被试应该可以正确排除 B 和 D 两个选择项，那么在该题上正确回答的可能性为 1/（4－2）＝0.5。最后再对每一题求和，可得及格线，若请若干专家同时评定，则可以这些专家所评定的及格线的平均值作为最终及格线。

（二）Angoff 方法

此法由 Angoff（1971）提出，由专家直接判断处于临界水平的被试在某测验的每一题目正确作答的可能性（记为 P_i），设每一题的满分为 F_i，则该测验的分数分界点（记为 λ）为：

$$\lambda = \sum F_i P_i$$

表 10.6 是利用 Angoff 方法确定测验分数分界点的实例，此例中假设测验欲测五个目标内容，记为①－⑤，且测验共有十道题目组成。

表 10.6　Angoff 方法示例

题号	目标内容	题目满分（F_i）	临界水平（P_i）	F_iP_i
1	①	2	.9	1.8
2	②	6	.7	4.2
3	②	6	.75	4.5
4	③	10	.8	8
5	②	6	.7	4.2
6	④	12	.65	7.8
7	④	12	.6	7.2
8	⑤	18	.55	9.9
9	③	10	.6	6
10	⑤	18	.5	9
		$\sum F_i = 100$	$\lambda = \sum F_iP_i = 62.6$	

此例及格线为62.6分，即在测验中得分在62.6分以上的被试评定为掌握者，反之则为非掌握者。同样，如果有多位专家同时评定，则以这些专家评定的平均及格线为测验最终及格线。

比较Nedelsky法和Angoff法，前者显然使专家的评定受到限制，若针对四择一选择题，专家评定的P_i值只可能为0.25，0.33，0.50和1.00，而Angoff法中的P_i则可在0—1.00之间任意取值，而且适宜于各种题型。因而Angoff法在实际运用中更受欢迎。

二、效标组预测法

（一）临界组法

由专家判定和选择一组正处于临界水平的被试，将测验施测于该组被试，计算他们在测验上的平均成绩，以体现测验的内容范围所要求的临界水平，因而可以视之为测验分数分界点的估计值。

采用这种方法的困难在于临界水平被试的选择与评定，一来要选出一定数目的临界水平被试必须先随机选取大量被试作为候选，二来对被试是否正处于临界水平很难找到客观而统一的标准，非常抽象而主观。因而此法的应用在实际中是受到一定限制的。

（二）对照组法

此法同样先采取专家判定的方法来选择被试，只是这里要事先确定两组被试，一组被明确判定为掌握组，另一组则被明

确判定为非掌握组，那些不太容易被判定为“掌握”或“非掌握”的被试一概剔除。对这样两组被试施测测验，可得到如图10.1所示的原始分数分布图。图中两条分布曲线的交叉点即为测验分数分界点（此图60分为测验分数分界点）。

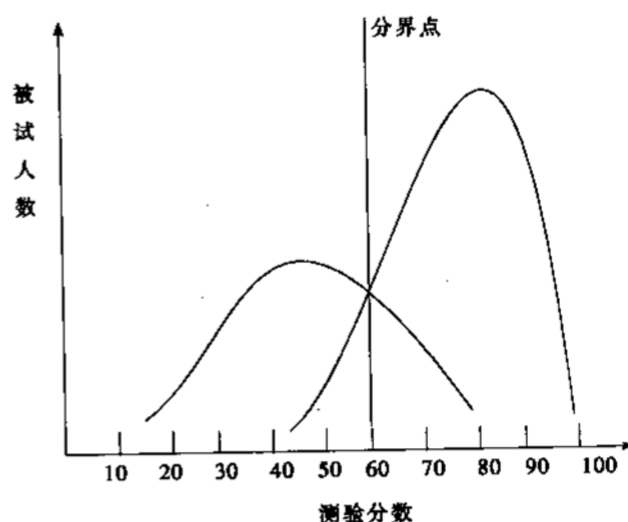


图10.1　对照组分数分界点标定法示意图

由于采用此法所标定的分数分界点在很大程度上会受到所取被试团体的影响，因此更合理的做法是选取若干对对照组，取每对对照组交叉点分数的平均值作为测验分数分界点。

由于判定被试为“掌握”或“非掌握”比判定其是否处于临界水平要容易很多，因而对照组法应用更广。也有人提出对照组的被试不用经过专家判断，直接取“已接受相应教学组”和“未接受相应教学组”，再以同样方法获得分界点值。这种方法虽然克服了专家判断所带来的主观性，但似乎带来更多的

问题：如何确定“已接受教学组”里的被试是否都已掌握了所教内容？又如何确定“未接受教学组”里的被试是否的确没有一个人掌握了相应内容？这些问题的存在使得由此得到的分界点的可靠程度反而招致更多的怀疑。

总之，在上述各种测验分数分界点的确定过程中，专家评定始终占据一定的位置，这就使得分界点带有一定的主观色彩，这也是人们对分界点的确定争论不休的主要原因之一。对此，一般应采取多位专家评定的方法，综合多位专家的意见，使分界点尽可能地合理与有效。

关于目标参照测验的分数解释，我们注重的是被试在所测内容领域的绝对掌握水平，并常根据实际需要而将被试在分数分界点上分为“合格”或“不合格”两类。但是，值得注意的是，对于目标参照测验的分数解释，有时并不排除同时采用一些常模参照测验的分数解释方法。目标参照测验和常模参照测验虽然是两种不同的测验模式，但它们并非互不相容，当我们既对被试在测验中的绝对水平感兴趣，同时又想了解被试的成绩在其所处团体中的相对位置时，该测验便兼具了目标参照测验和常模参照测验的性质，对其测验分数，则既可以采取上述方法来解释，同时又可以采取常模参照测验的方法给出其百分等级等有关相对位置的信息。

练习与思考

1. 请根据表 10.3 中专家评定结果的平均数对表中 14 道题目的内容效度进行评定，并决定题目取舍。

2. 目标参照测验的题目区分度可以通过哪些方法来确定？

3. 目标参照测验的信度与效度为何不适宜以相关系数为指标？

4. 试比较目标参照测验与常模参照测验的异同。

5. 试分析内容范围的界定在目标参照测验的标准化过程中的重要性。

6. 试分析分数分界点在目标参照测验的质量评估中的作用。

7. 试析题目内容效度与目标参照测验的内容效度间的联系。

第十一章　学绩测验

本章提要：

- 学绩测验的性质、作用与分类
- 标准化学绩测验的性能与编制
- 题库与题库建设
- 史坦福成就测验与关键数学算术诊断测验
- 教师自编课堂测验

第一节 学绩测验概述

前述各章我们向读者详细介绍了心理与教育测量的一系列基本原理和各种计量分析技术。从本章起我们将根据所测心理品质的不同向读者具体介绍几种重要的心理与教育测验。

首先介绍学绩测验。学绩测验是心理教育测验中发展比较早的一种测验。据史书记载：早在我国的西周就初步建立了学校教育制度，那时国学中的大学就已设置了定期的学业考试。学校考试沿袭至汉朝时，太学中已经订有严格的考试制度，武帝时规定一年考一次，到东汉恒帝时改为“二岁一试”。太学考试的方法有“口试”、“策试”、“射策”等3种，通过考试者毕业时按成绩授予不同的官职。我国历史上沿袭了1300年之久的科举考试是当时世界上规模最大、影响也最大的，由国家组织的学绩测验。在西方，古代有名的教育家、哲学家苏格拉底在授课时就采用了口试方法。中世纪的欧洲，各大学均以口试作为毕业成绩考核的方法。18世纪末19世纪初，欧美各国也开始实行用学绩测验考核官吏的文官考试制度。

学绩测验源远流长、它基本上是与学校教育同步产生的。学绩测验应教育的需要而产生，是服务于教育的一种重要手段，也是教育过程中的一个重要环节。学绩测验在当今世界上应该是应用最为广泛、最为频繁的心理与教育测验了：各级各类学校的各种学科测验、招生考试，各级各类行政企事业单位的招干、招工考试，各行各业的上岗、晋职考试都属于学绩测验的范围。当今社会一个人从求学到退休，恐怕很少不经历过

数十次乃至数百次的学绩测验考试的。有的人今天当主试，明天又成了被试，后天可能又是主试了。因此，学绩测验也是人们最为熟悉的一种心理与教育测验。测量学者历来就非常重视对学绩测验的研究，在学绩测验的原理和编制技术的研究上都取得了很大的成功。由于学绩测验的内容和形式都非常丰富，社会对于学绩测验的需求也各种各样，因此我们更应该重视对学绩测验理论的研究。

一、学绩测验的性质

"学绩"一词通常是指个体经过对某种知识或技术的学习或训练之后所取得的"成绩"，一般表现为个体心理品质在知识、技能或某种能力方面的增加和提高，是个体认识性心理品质的发展。无论个体学习的知识或技术的内容是什么，也无论个体采取的学习或训练形式是什么，我们都会对个体的知识增长量和技术能力增长量感兴趣，都希望能对个体的知识、技能增长量或是当前的知识、技能发展水平进行数量化的测定，这就是学绩测验的目的。

学绩测验是对个体在一个阶段的学习或训练之后知识、技能的发展水平的测定。学绩测验与一般的心理测验不同。一般的心理测验所测的往往是为个体各种经验积累以后的一般心理发展水平，有的甚至要排除那些"专门"的学习或训练的影响而测个体"稳定不变"的心理品质。学绩测验则相反，它更希望测量个体通过一次或一个时期的学习训练之后，这种专门的知识和技能的发展水平。理论上甚至认为，学绩测验所测之内容不经过专门的学习和训练，其测值应该几乎为零。若不是这

样，则所编制的学绩测验是低质量的、不成功的。

学绩测验与能力测验一样在测量学中属于最佳行为测验。最佳行为测验施测时要求被试调动他所学的一切知识、所具备的一切技术和能力，对所有试题给出最佳答案或最佳操作。从这个角度看，主试与被试的目的是完全一致的，都是为了测出被试的最高发展水平。因此，编制学绩测验对于主试来说就是要设计出与被试认知特质紧密相关的试题并组拼成试卷，通过施测、评阅将被试的认知发展水平与一个数字系统中的某个确定值相对应，以便区别被试的水平差异。与典型行为测验不一样，学绩测验不用担心被试在测验上故意掩盖自己的行为水平，相反却担心所编测验达不到诱发被试发挥出最高水平的目的。当然学绩测验也要防止被试用猜题、押题等“针对性”的学习和训练获取“好”成绩的现象。

学绩测验所测为认知性心理品质。认知性心理品质的优劣表现在两个方面：一方面是认知内容的多寡，一方面是认知能力的高低也就是我们通常所说的知识与能力两个方面。学绩测验发展至今，已经比较重视开发测知识与测能力并重的测验，纯测知识的测验已不受人们的欢迎。但是学绩测验与一般的能力测验又不相同。能力测验往往更强调所测为“一般能力”，而排除知识，特别是“专门”知识的影响。尽管能力测验实际上也要通过测被试对知识的理解、应用等操作行为而实现，但其重心是在能力。而学绩测验却是知识与能力并重，即使测能力，也是测对所学专门知识的理解、应用等能力。我们不能把学绩测验编制成一般能力测验。

学绩测验通常用于对个体经学习、训练之后学习成绩的鉴定和诊断，有时也用来预测被试在今后的学习或工作中的成就，但是它与一般的性向测验又有不同。学绩测验是针对一有计划的学习或训练之后的成绩的测定而设计的，所测认知能力

较具特殊性，即使用来预测，也是由于所预测的学习或工作与这种特殊的学习或训练紧密相关。

一般性向测验往往开始于某种专门的学习或训练之前，希望测试被试在以往的生活经验中获得了多少与这专门学习或训练有关的能力，以预测被试在即将开始的学习或训练中的成就。性向测验所测认知能力较具广泛性，有时还带有情感因素，其根本目的是要为被试能不能参加这种专门学习或训练提供依据。

二、学绩测验的作用

学绩测验的作用非常明显，学校使用学绩测验鉴定学生的学业成绩。学生经过一个阶段的学习，到底获得了多少知识，提高了多大的能力，可以通过学绩测验进行测定。学绩测验的结果反馈给学生，学生可以总结学习经验，纠正不足，利于学生进一步学习；学绩测验结果反馈给教师，教师可以总结教学经验，利于教师进一步改进教学。学校还使用学绩测验甄别学习困难儿童，诊断学生学习困难的原因，以便及时制定和采取补救措施，帮助学生全面掌握所学知识，全面提高专业能力。学校还应用学绩测验辅助教学管理。升学、毕业、升级、留级、划分班级组别都需要学绩测验的信息。现代社会的人事管理也应用学绩测验。人员取录、晋职提级都可以利用学绩测验，以测验成绩作为重要的取舍依据。没有学绩测验提供准确的信息，教育管理会陷人混乱，人材使用就会陷于盲目和造成浪费，人力资源难以得到合理配置。教育科学研究也需要学绩测验。教育科研工作者利用学绩测验信息评价教育决策、优选

教育方案，为教育的改革和发展作出独特的贡献。

三、学绩测验的分类

（一）按测验的编制方法分

按测验的编制方法可以把学绩测验分为教师自编课堂测验和标准化学绩测验两大类。教师自编课堂测验由教师根据自身经验编制，所测内容可多可少，时间可短可长，主要施用于自己的学生，紧密结合教材和教学实际，形式活泼多变，可用来考查学生学习情况，也可用来检查教师教学质量，甚至可以用来预测学生未来成就。但教师自编课堂测验应用范围较小，不能在大范围内对学生进行比较，大多数教师只有专业知识而没有测量学知识，仅凭个人经验命题，随意性大，效果往往不理想。标准化学绩测验由测量学专家与学科教师按测量学基本原理编制，有一定的质量指标做保证，能提供常模作比较，客观性强，可用于大规模正规测试。但是标准化学绩测验编制费时费力，灵活性和针对性均不强。因此，学校教育中使用更多的还是教师自编课堂测验。

（二）按测验内容分

按测验的内容对学绩测验进行分类通常是以材料内容所涉及的学科分。有单科测验如语文测验、数学测验、生物测验等等，也有多科测验。多科测验常以组合测验形式出现，比如某一个年级的综合测验，包括几个学科分测验。多科测验用以评价学生的总体水平。

按测验内容分类也有以内容量的多寡分的，如单元测验、

总测验等。

（三）按测验的用途分

按测验的用途可把学绩测验分为考查性测验和诊断性测验两大类。考查性测验主要用于对学生学习结果的鉴定。学校的单元测验、期中测验、学科结业测验，社会的招生考试、招工考试、提职晋级考试都是考查性学绩测验。诊断测验主要用来测查学生在各个具体教学内容、教学目标上学习的长处和弱点，分析学生学习困难的原因，并提出相应补救措施。诊断测验多以单科内容为测验材料，编制时都是从非常细微的地方入手，以获取详细的信息。诊断测验在对学习障碍儿童、学习缓慢儿童的鉴别评定上也具有较高的实用价值。

（四）按测验评分的参照系分

按所编测验评分系统的参照系不同可把学绩测验分成常模参照性测验和目标参照性测验两大类。常模参照性学绩测验以学生伙伴总体为参照系，以学生在伙伴中的相对位置评价学生的学习成就。目标参照性测验以教材和大纲为参照系，以学生有否达到教材与教学大纲规定的教学目标来评价学生的学习成就。常模参照性学绩测验易于横向比较，常用于选拔性目的的测量；目标参照性测验以教学目标为准，常用于鉴定学生的合格与否。

（五）按测验的题型分

学绩测验可使用的试题大致可分为定向反应型和自由反应型两大类，习惯上又分称为客观型试题和论文式试题，因此也有把学绩测验分为客观测验和论文式测验两类的。两大类题型的性质与功能在第三章已作详细分析，并证明两类题型间有互

补作用。因此我们建议，不是有特殊需要，不要使用单一式的试题组成学绩测验，还是以两大类题型配合使用为佳，至于测验中两类题型之比，可根据情况作适当调整。

学绩测验还可根据一次施测的被试多少分为团体测验与个别测验两种；还可根据被试反应的行为方式分为口试、笔试和实验操作等3种。操作测验在技术、技能测验中应用较多。特别值得提及的是一种新的测验形式，即在计算机上进行的测验。这种测验不是对被试的计算机知识或操作水平进行测定，而是利用计算机进行其它学科的测验，故称其为计算机化测验。计算机化测验在形式上把命题、组卷、出示试题、考生作答、评分等一系列的测验管理工作集中在一起一次完成，可节省大量的人力、物力，而且评分客观、公正，保密性能好。若配以辅助设备，有些操作性测验也可以在计算机管理下完成。计算机化测验是测验科学与计算机技术相结合的产物，表现出众多的优良性能，受到社会的欢迎。

第二节　标准化学绩测验

在心理与教育测量学原理指导下，遵循一定的程序所编制的各方面质量都达到规定标准的学绩测验，称为标准化学绩测验。长期以来人们对于仅凭主观经验编制的各种学绩测验的批评一直都是比较激烈的。人们认为这类测验的测验目标不明确、测验内容不统一、测验的标准不一致、测验的结果不精确，人们希望能编制出目标明确、标准一致、精确性高的新测验。标准化测验就是应这一要求而产生的。标准化测验在许多

国家应用比较广泛，在我国也越来越受到重视，我国高考的标准化试验研究就是国内影响最大的编制标准化学绩测验的尝试。本节重点就是对标准化学绩测验作一深入的介绍。

一、标准化学绩测验的基本要求

一份测验能称为标准化测验，最起码要符合以下几方面的要求。

第一是命题组卷标准化。标准化试卷的所有试题都是经过精心编制的，试题测量目标明确，语词意义清晰，试题难度、区分度达到规定标准。标准化试卷全卷结构与测量的目标系统一致，知识覆盖面宽，题型比例恰当，题量适度，试题难度分布符合规定要求，试卷的信度、效度都达到规定标准。标准化测验通常还要备有等值复份。命题组卷标准化的目的是提供一份高质量的测验试卷。

第二是施测标准化。标准化测验必须在统一标准的环境下施测。统一标准环境包括测验场所的标准统一、测验时间统一、测验的指导语统一、提供给考生的测验材料统一、材料出示的顺序统一。有的标准化测验还包括考前给考生提供统一的考试大纲，考后提供统一的标准答案。测验场所的标准统一既包括场所外环境又包括场所内环境。外环境包括噪音指数、温度指数、湿度指数等等；内环境包括空间、面积大小，主试被试人数，主试被试位置及被试与被试的位置间隔等等。施测标准化的目的是给被试提供一个公平、优良的施测环境。

第三是评分标准化。标准化测验的评分在测验编制的同时就要制定好标准答案和评分规则。标准答案要正确、规范，最

好是唯一。评分规则应尽量细致、客观，最好是没有伸缩性。如果人工阅卷，要求阅卷者有高度的责任心，有较高的业务水平和较强的评判能力，要能够尽量维护评分标准的一致性。大规模阅卷还必须先进行阅卷培训，统一认识，统一标准。阅卷时采用流水作业法，并加强自查和复查。标准化测验能够应用机械阅卷的应尽量采用机械阅卷，以便提高工作效率，降低阅卷误差。评分标准化的目的是提高测验评分的精度。

第四是测验分数解释标准化。对于常模参照性测验，其意是编制测验时必须搜集常模样本，编制好测验常模。测验常模供被试查阅以便被试准确评价自己的测验成绩，测验也可直接将被试的常模分数通知被试。对于目标参照性测验其意是在编制测验时要认真研究教材和教学大纲，分析合格标准的确切含义，通过调整试卷难度结构准确划定合格分数线，测验后按被试卷面得分判断他是否达到规定的教学要求。

要编制出符合要求的标准化学绩测验，必须由测量学者和各学科专家共同合作，根据心理与教育测量学的基本原理，结合学科特点，遵循一定的编制程序，应用计量分析手段，精心设计、精心命题、科学组拼并经过反复试验，才能获得成功。

二、标准化学绩测验的编制

在前述各章，我们已将测验编制的基本原理及各种计量分析技术逐一作了介绍。编制标准化学绩测验只要在附加若干标准化条件的情况下，将这些原理和技术应用于学绩测验的编制。下面我们结合学绩测验的特点将标准化学绩测验编制方法分步骤介绍如下。

（一）确定测验目的，选定测验编制的方法

编制标准化学绩测验的首要问题是要确定测验目的，即要解决为什么而测，测谁和测什么的问题。首先要明确为什么而测，如果是为了考查学生学习情况，则要编制考查性学绩测验；如果是为了诊断学生的学习困难，就要编制诊断性测验。同是考查性测验还需分清是以比较考生优劣为目的还是以鉴别考生在学科学习上是否达到规定标准为目的。若是为了前者，则要编制常模参照性测验，若是后者则要编制目标参照性测验。测验的目的不同，测验的性质也不同，所依据的测验编制原理也不相同，编制的方法也就不同，是不能混淆的。确定测验编制的目的还要明确测什么的问题。测什么的问题包括所测是什么学科，是单科还是多科，是部分还是全部。是哪门学科的测验应该有哪门学科的专家参与，所编测验也应有那门学科的特色。明确测谁的问题也很重要。测谁的问题包括年龄年级特征，文化背景特征等问题。施测于不同对象的测验应该有不同的编制特征，具体可表现在题型选择、难度层次、教学目标层次等多方面的不同，甚至还会有城乡差异、民族差异、宗教差异、语言文字差异的表现。只有把以上问题一一明确了，才能够说测验的目的明确了，才能准确选择试卷编制的方法。

（二）分析测量目标，拟定测验编制计划

分析测量目标是制定测验编制计划最重要，也是最困难的一项工作。分析测量目标要应用到学科专业知识、心理学与教育学理论知识，还要有较丰富的教学实践经验。通常编制测验有一个总的测验目标，但是总目标往往太抽象、太笼统。因此，要根据认知理论将总目标分解成系统的认知目标体系。这项分解工作既要符合总目标的原则，又要尽量细致、明确，并

且要尽量可操作化，只有这样才能供实际编题使用。这个认知目标体系也就是通常所说的教学目标分类体系，它是供编制测验使用的，因此要测量“专业”化一些。国际上比较流行的美国教育心理学家布鲁姆的教学目标分类体系，也常常用来作为测量目标分类体系。但是不顾及具体情况而全套照搬是不恰当的，使用时应针对不同对象、不同学科作出适当的调整。一旦建成了针对具体测验的测验目标分类体系，就可以着手制定测验编制计划了。首先是编制一份测验双向细目表，将测验的内容分类与测验的目标分类共列于内，定出各个分类组合在测验中的占分比例。其次应该确定的是使用题型的种类及各种题型的占分比，以及全卷试题的难度分布。各项比例确定之后还应把全卷的结构统筹分划，定稿成正式的测验编制计划。

（三）编题征题与选题组卷

试题是测验的主要组成部分，是测验质量高低的主要体现。试题的来源可以组织学科命题教师自己编写，也可向社会征集。无论是自己编写还是向社会征集，试题分布都必须符合测验编制计划所定的测验结构，特别是要严格按照测验双向细目表的要求编写试题，不要编写细目表中未列的试题。编题时还要求命题者同时提供参考答案和评分标准供审题参考。有了试题只是完成了第一步，接下来要进行试题筛选。试题筛选有两方面的工作要做。一是对试题的文字内容进行审查，内容是否科学、逻辑是否严谨、文字表述是否准确清楚等等，都是审查的内容。其次是进行试测。通过试测获取试题的难度、区分度指标。对那些区分度偏低或难度不合要求的试题进行修改，或干脆淘汰。经过筛选留下一批高质量的试题供组卷使用。组卷时应严格按测验计划进行，内容分类比、目标分类比、题型比、难度比都要符合预定要求。对于人选试题还要进行编排。

编排顺序通常是全卷按题型分类，题型顺序按先简后繁排列，同题型内按难易顺序编排。试卷编排完毕还应写好测验指导语，对于新异题型要编写解题范例。特别要注意的是标准化测验应同时编制等值复份，等值复份的编制要求与原本编制是完全一样的，最好是在编完以后再随机确定哪是正本、哪是复份。

（四）调查测验质量参数，编制测验常模

标准化测验要得以发行使用，必须提供测验质量参数，包括测验的信度、效度等指标。若是目标参照性测验还必须提供合格分数线，有的还要提供误判概率。常模参照性测验还必须提供测验常模。测验质量参数和测验常模都要通过取样测试。选择测试样本（包括前面试题测试求取试题参数的样本）要注意保证样本对总体有充分代表性，这就需要样本有一定的容量。如果被试总体层次结构复杂，还必须采用分层随机抽样方法获取测试样本。样本有充分的代表性就能保证所获参数真正反映试卷的质量，也使得编制的常模准确反映被试的总体状况。参数计算方法和常模编制方法前面章节已有介绍，此处不再赘述。如果所编测验质量参数达不到要求，说明试卷质量还不符合要求，编制人员必须仔细分析原因：若是试题质量不高，如区分度不高、难度不合要求、所测教学目标不准确等，则应撤换试题；若是试卷结构不合理则要修订测验计划。常模编制时，若被试层次结构太复杂，层间差异很大，可能还要考虑编制分常模，以供不同层次对象使用。

（五）编写测验指导书，正式出版发行

测验质量达到规定要求，常模也已编好，测验的编制进入最后阶段，那就是编写测验指导书，连同编排好的试卷（包括

答卷纸）一起正式印刷发行。测验指导书内容包括测验目的、适用对象和范围、测验操作要求、测验质量参数、标准答案、评分规则等项目。测验常模可以附印在指导书后，也可以单独印刷。当然正式发行还需有负责机构的批准。

三、国外常用标准化学绩测验简介

（一）史坦福成就测验

史坦福成就测验（Stanford Achievement Test）属于综合性学绩考查测验，也是一种供团体使用的常模参照性测验，使用历史比较长，初版于 1923 年，中间经过多次修订，颇受社会好评。该学绩测验是一种组合式测验，纵向可分成 6 个不同的级别水平，适用于 1～9 年级学生。具体级别划分见附表 11.1。横向包括 11 个方面的科目内容，分别为词汇、阅读理解、拼字、听理解、词汇学习技能、语言、数学概念、数学计算、数学应用、社会科学常识和自然科学常识，基本覆盖了美国中小学生所有的学习内容。这些科目内容又分别组合成不同的分测验供实际使用。但是在不同级别上，科目和分测验数又有不同。初级 1 有 5 个分测验，初级 2 也有 5 个分测验，但增加了部分科目。其他级别均有 7 个分测验，其中社会科学常识和自然科学常识分测验在中高年级中是通用的。

表 11.1　级别划分表

级别名称	适用年级
初级 1	1.5—2.9
初级 2	2.5—3.9
初级 3	3.5—4.9
中级 1	4.5—5.9
中级 2	5.5—7.9
高级	7.0—9.9

史坦福成就测验还有两个配套测验。一个叫史坦福早期学校成就测验，一个叫史坦福学业技能测验。前者适用于幼儿园和一年级学生，后者适用于八到十三年级学生（即初二到大一年级）。当然它们各自包括的科目是不同的。史坦福成就测验还配有练习测验。练习测验提前两天提供给被试练习，协助被试熟悉测验特点。正式测验各级别所用时间在 3 小时 35 分钟到 5 小时 15 分钟之间不等，因而往往分在几天内完成。

该测验现行版本提供两套常模，学年初常模和学年末常模。常模样本分别包括来自 300 多个学区的 25 万秋季测试学生和 20 万春季测试学生，采用分层随机抽样方法获得。该测验使用的导出分数有 5 种形式：百分等级、标准九分数、年级当量、量表分数和正态曲线当量。据报告，史坦福成就测验各分测验的信度均在 0.80 以上，总测验的信度高于分测验信度，高级别测验信度高于低级别测验信度，同级别分测验间的相关均在中等程度以上。测验的内容效度和结构效度均获得符合要求的有力证明。

(二) 关键数学算术诊断测验

关键数学算术诊断测验（Key Math Diagnostic Arithmetic Test）初版于1971年，适用于学前儿童直至小学六年级的学生。测验分成内容、运算和应用三大块。内容块有3个分测验：数学、分数、几何与符号，主要测量基本的数学概念和知识。运算块有6个分测验：加法、减法、乘法、除法、心算和数字推理。应用块有5个分测验：文字题、补充、金钱、测量、时间。这是一个个别测验，全部测完需30到40分钟的时间。关键数学诊断测验在4个层次上对被试进行数学技能诊断。第一个层次是总体水平诊断，指出被试在同年级伙伴中的位置。第二个层次是分块水平诊断，比较被试在内容、运算和应用三块上的强弱。第三个层次是分测验水平诊断，比较被试在14个分测验上的高低差异。第四个层次为项目水平诊断，直接指出被试在各个项目所代表的内容和教学目标上的理解程度。每个层次的分析都备有侧面图，诊断结论显得非常清楚。该测验还别具匠心，备有与各题目相关联的行为目标清单，供设计教学补救计划参考。据报告，该测验的常模样本包括了1222个幼儿园到七年级的学生，来自美国8个州21个学区。该测验的总分信度的中值为0.96。有研究者还报告该测验的部分分测验的并存效度分布在0.38到0.63之间。

在我国，二三十年代时曾掀起过编制标准化学绩测验的热潮，解放后台湾地区的学者在测验编制方面也一直做出努力。大陆上正式起步比较晚一些，用于校内的学绩测验较有影响的还比较少见。现在国内标准化学绩测验研究的关注中心是高考的标准化试验，我们将在稍后专门介绍这方面情况。

四、标准化学绩测验的题库建设

一些大规模的标准化学绩测验应用范围广，施测周期短，对试卷的需要比较频繁。每次都临时编题组卷，耗费大而效率又不高。解决这个问题的一个有效办法就是建设一个题库。应用题库组拼标准化学绩测验的试卷，具有经济、高效、而且保密性强的特点。国外在第二次大战后开始研究心理与教育测验的题库建设，70年代是题库建设和计算机组卷技术发展最快的时期。我国的题库建设近十多年来也受到了各方面的重视，报刊上常见某科某种题库建成的报告。但统观这些题库，质量和性能的差异很大，差的充其量只能视其为一个“题集”而不是题库。所以有必要加强对题库建设的研究和宣传，鼓励多建高质量的题库。

通常认为，一个高质量的题库应具备以下几个方面的优良性质：

（1）植基于一种科学的测量理论。

（2）贮备有一定数量的试题，所有试题品质优良，技术参数完备。

（3）题库内部结构层次清楚、分类严谨，试题检索方便。

（4）题库管理方便、可控性强、易于维护更新。

（5）保密性强。

更理想的题库还应实现计算机管理且应配备计算机自动组卷程序，充分开发题库功能。学绩测验比其他任何心理与教育测验的应用都更广泛，更频繁，也更需要保密。因此建设一个高质量的题库是大规模学绩测验维持测验高效、质量稳定、标

准一致的必要条件。下面我们就学绩测验题库的建设方法作一些介绍。

首先是选定一种构建题库的测验指导理论。没有科学的测验理论作为指导，就难以合理规划题库结构，难以科学制定选题原则，难以建成高质量的题库，也不能按测验原理从题库中有目的地选择试题组拼出质量符合要求的标准化试卷。通常可选的测验指导理论有经典测验理论和项目反应理论两种，经典理论就是本教材重点介绍的内容，项目反应理论我们将在最末一章向读者介绍。

其次是设计题库结构。题库结构应根据选定的测验理论模型进行设计，主要应包括以下几方面的内容：

(1) 首先确定题库中试题所应用参数的个数，各种参数使用名称。通常有试题内容、教学目标、题型、难度、区分度等可选参数。实际用多少、用哪些应视建库者的目的、技术和人财物三方面情况而定。

(2) 确定全库试题的内容范围及内容层次详目。内容层次分得越细越好。

(3) 确定全库试题教学目标层次详目。各目标层应尽量用操作性语言叙述。

(4) 确定全库试题的题型种类数及具体题型。题型种类数不宜太少，一般在 10 种左右。不必刻意追求题型的新颖性，但要保证所使用题型应有较好的性能。

(5) 确定全库试题难度等级的划分。这个划分可以粗放一点，用难、中、易三等也可。目的是宏观控制一下全库试题难度水平的比例。

(6) 确定题库总题量及在各参数层次上的分题量。最理想的应该定出每一参数组合的具体题量，即定出××内容、××教学目标、××题型、×等难度的题量为多少，以便建库时有

目的有计划地编征试题。

建库的第三步工作是编题、征题、试测、分析、筛选、编码入库等一系列具体操作。前几项具体操作与标准化学绩测验中所叙要求类似，只是题量要满足计划的要求。编码入库时除将试题本身的文字内容送入题库，还应将试题的各项参数指标，包括区分度、难度等数值，以及试题的标准答案、评分标准一起存入题库。

完成了以上三步工作，题库就初步建成，可供使用了。如果是计算机贮存管理的题库，还可以开发各种组卷软件，自动组拼用于各种目的的学绩测验试卷。如果不是计算机化题库，则应注意题库的保密问题，应制定专门的题库保管和启用规则。

题库建成后还应不断地进行维修更新，不应成为一个固定不变的“死”库。日常的维修更新工作主要是定期复审、复测，淘汰或修改那些内容陈旧的和性能退化的试题。同时补充内容新、性能好的新试题。用这种“新陈代谢”的方法，防止题库老化，延长题库的使用寿命。建设一个题库投入比较大，但题库建成后使用的效率高、效果好，因此值得测量工作者一试。

五、我国高考的标准化试验

我国的高考从规模与影响来说，在国内都是首屈一指的。每年一度的高考牵动着成百上千万学生和家长的心。国家每年都要投入大量的人财物力主办这么一次规模巨大，政策性、技术性都很强的国家考试，为高校选拔几十万大学新生。由于国

家重视，也由于考试工作人员的努力工作，高考也是国内最有信誉、最有权威的考试。但是，考试标准不稳定，考试结果误差较大，选拔线不“公平”等弊病却一直存在，这在一定程度上影响了高考的声誉。为了维护高考的权威，提高考试质量，更准确地选拔人才，我国开始了高考标准化的试验研究。从根本上来说，我国高考标准化试验的目的就是要应用现代心理与教育测量学的原理，对传统高考进行科学化改造，努力提高考试命题和考试管理的水平，努力提高考试的信度和效度，逐步达到标准化考试的水准和要求。

我国高考的标准化试验在 80 年代初就开始酝酿，1981 年到 1985 年国家教委（原教育部）学生司就多次召开过有关高考改革的研讨会。1985 年开始，受国家教委委托，广东省开始高考的标准化试验，上海开始了高中毕业会考试验。广东省的此项任务是对考试的命题、施测、阅卷及分数解释进行标准化试验研究。试验从 1985 年的英语、数学两科，逐步发展到 1988 年的 5 科；英语科试验省份率先增加，1988 年发展到 17 个省市；1992 年开始，全国普遍推行主、客观题分卷印刷，客观题实行机器阅卷两项措施。上海的试验以建立高中毕业会考制开始，最终达到会考、高考配套接轨，高考科目改组的目的。1994 年会考制在全国普遍实行，新高考制也逐步实行，考试科目实行 3+2 制，即公共必考 3 科：语文、数学和外语，同时文科另加政治和历史，理科另加物理和化学。

在阅卷评分方面，受国家教委委托，江西省于 1989 年首先在全国统考卷中进行作文评分改革试验。随后试验逐步扩大到河北、河南等省，1994 年分项分等评分方法已在全国大部分省份推广。当前正在试验研究并准备在全国推广的是高考分数的标准化制度。经过短短 10 年的试验研究，我国高考标准化的程度从各个方面都有了很大的提高，获得了公众的好评。

高考标准化试验的初步成功，为我国考试事业的发展打下了良好的基础。

必须要强调的是，我国高考的标准化并没有全盘照搬国外的理论。我国高考的标准化是在坚持和继承我国传统考试的优良特色的基础上实现的，主要表现在我们坚持主客观题并用，坚持考核高层次教学目标，特别是在主观题考试的标准化研究上作出了努力，这是国外标准化学绩测验所不及的。

第三节　教师自编课堂测验

标准化学绩测验的质量高、误差小，能在大规模测验中客观准确地完成对被试的测定和评价，是很理想的测验形式。但是标准化学绩测验的编制技术要求高、投入大，特别是要考虑到全面，所以灵活性较差。学校教育对测验的要求千差万别，没有哪一种标准化学绩测验能够同时满足各种要求。因此，学校教学中使用的大多数测验还得依靠教师自己设计、编制、使用、评分。一定意义上说，学校可以没有标准化学绩测验而不能没有教师自编课堂测验。但是我们应该看到，目前教师自编测验的质量是不能令人满意的。必须鼓励教师学习一点测量学的理论和技术，努力提高自编测验的质量，使自编测验更好地为教学服务。

一、教师自编课堂测验的特点

（一）测验形式灵活多变，与测验目的完全一致

教师自编课堂测验完全由教师根据测验目的自己确定测验的时间、地点、内容和形式。测验时间短的只有几分钟，十几分钟，长的几十分钟，甚至上百分钟。测验形式可以笔试，也可以口试，甚至实验操作，有的还可以混合使用。测验内容可以是一门课程、一册教材，也可以是一个单元、一篇课文。测验可以停课举行，也可以穿插于授课之间。测验对象可以是一个学生，也可以是一个小组、一个班级，也有几个班联考的。

（二）测验内容与教材内容高度一致

教师自编课堂测验完全依据教师自己所用教材编写，不必顾及其它教材的内容和形式。因此，教师自编测验内容与教材内容可以高度一致，可以考出所选教材的特色，特别是对于选用地区教材、乡土化教材的学校更为贴切实用。使用教师自编测验为教师自主安排教学内容和进度，教出特色提供了方便。

（三）测验难度切合学生的实际水平

标准化学绩测验的对象分布范围广，所编试卷只能以全体被试的平均水平作为难度的参考水平。但是教学的地区差异、学校差异乃至班级差异都是客观存在的，用一个平均难度的测验去面向全体被试，虽然保证了被试大范围内的可比性，但在更多的地方却显得难度不是很恰当。对于水平偏低的被试，测验显得难了，对于水平偏高的被试，测验又显得容易了，这都

不利于调动学生的学习积极性。而教师自编课堂测验由教师自己编写，可以针对自己学生的实际水平出题，因此就不会出现难度不恰当的现象。学生经过测验均能从中获得针对性很强的评价信息，找到自己努力的方向。

（四）测验编制简易快速

教师自编课堂测验是在教师对学生、对教材、对教学要求都非常熟悉的情况下进行的，若他又注意积累了以往的教学和命题经验，那么编制一份测验是比较容易的。况且教师自编课堂测验不需要试测，不需要收集信度、效度数据，更不需要什么常模，因此编制花费的时间也不需要很多。大多数教师自编测验编制简易快速，可说是学绩测验中的一支轻骑兵，只要教学需要，它就能快速地实现施测，这是任何标准化学绩测验无法做到的。

二、教师自编课堂测验的步骤与方法

历来教师自编课堂测验都是由教师自己设计、自行编制的。但是，教师的水平有高低，经验有多寡，工作态度有好坏，在编制测验时自觉防止主观化、片面化的意识有强弱，因此，所编制测验的质量有高低。我们认为应该鼓励教师学习一点心理与教育测量学知识，应该按照测验编制的科学原理去规范教师的测验编制行为，去提高教师的测验编制技术，努力提高教师自编测验的质量。下面根据这个观点，介绍教师自编课堂测验应遵循的一般步骤与方法。

（一）审查测验目的

教师自编测验往往自认为对测验的目的是很清楚的，因此不会过多地思考。这对测验编制是不利的，教师应该认真审查自己编制测验的目的。单知道测验对象是谁不够，还应深入了解这些对象的学习水平和特点；单知道测验内容是什么还不够，应该明确具体的教学要求和教学目标是什么；单确定是学绩测验还不够，还应明确是什么性质的学绩测验，是常模参照测验，还是目标参照测验，是考查测验还是诊断测验，是速度测验还是难度测验。只有把这些有关测验目的的细节都考虑清楚了，才可能编制出真正符合测验目的的测验。关于测验性质的区分是很重要的，学校使用的学科结业测验应该是目标参照性测验，如果编成常模参照测验，就没有客观稳定的合格标准，很可能造成对学生的误判。把不合格学生误判为合格的，是对学生不负责任，把合格学生误判为不合格的，就会挫伤学生的学习积极性，这样都会产生不良后果。所以教师在动手编制测验前应认真审查自己的测验目的，理清自己的思路和各种关系。

（二）制定测验编制计划

教师自编课堂测验也应该有一个详细的测验编制计划。编制测验既要在试卷结构上全面合理安排，又要在一个个测题上深入细致地研究，这就要求思维既要有广度，又要有深度。人的思维毕竟有限，能同时照顾到广度又照顾到深度的并不多。制定一个详细的测验编制计划，就是促使教师首先在整体角度认真考虑，计划完成后，则可以按计划要求，在各个测试点上深入研究，这样既保证了广度又达到了深度。测验编制计划的具体内容可参考前一节有关部分，值得提醒的是对于教学目标

的分类，教师应紧密结合测验实际，制定详细的分类体系，切忌生搬硬套。

（三）命题与组卷

命题与组卷的首要问题是教师应该严格按照测验计划进行，特别是要严格按照测验双向细目表规定的教学目标编制出符合要求的试题，这对一般教师来说是件比较困难的工作。因为教师对于教学目标的理解往往不一致，即使理解准确了，要编出恰能测定这些目标的试题又是一个难题。所以教师应该结合专业，学习一点心理与教育学知识，并且注意积累经验，在命题中逐步提高命题技术。教师命题还有三条要求：一是提前，二是超量，三是审查。提前是指早一些把题命好。因为有些试题在编制时觉得是得意之作，隔了一个阶段再来看却会发现意义不明确，甚至模棱两可。超量是指多命一些题，然后从中优中选优，保证试卷质量。审查的含意有两重：一是自己认真审查，包括自己解答一遍；其二是请别的老师审查一下，做一做。有的题自己形成思维定势，可别人一看马上就发现有歧意或错误。关于组卷，除前面所说要严格按测验计划选用试题外，还应考虑试题的编排顺序。这方面的要求基本也可参照本章前节的有关内容。另外，教师也应过问试卷的编辑和印刷。编印的一般要求是清晰、正确、有条理，要留足学生作答空区，要注意不要把一道试题分印在两页上，影响学生解题。

三、教师自编课堂测验应注意的问题

教师自编课堂测验在编制与应用中还有以下几个方面的问

题需要注意：

（一）教师要深入研究教材，深入调查学生

由于是自己教、自己考，教师认为自己已经很熟悉教材，也很熟悉学生，所以就不再去认真研究教材、调查学生，凭自己的经验命题。这样容易形成所测内容往往只涉及自己熟悉的或自己注重的内容，考试对象只以自己熟悉的或亲近的学生为代表，这样造成所命试卷内容不全面，甚至缺漏某一个或几个重要方面的结果，或者造成不能适合大多数学生的后果。所以命题时，教师还是应该重新深入研究教材，深入调查自己的学生。深入研究教材时还应注意在命题时以测教材的基本原理、基础知识为主，内容不宜太锁碎，还要贯彻既测知识、又测能力的原则，不能把课堂测验变成纯知识的测验。

（二）要维护准确稳定的合格标准

教师自编测验多数是目标参照性测验，目标参照性测验编制的一个重要原则是要维护准确稳定的合格标准。所谓准确，就是要保证凡被评为合格的学生确实都是基本掌握了本门学科内容，达到教学基本要求的人；凡评为不合格的学生确实是未达到学科基本要求的人，不能有过多的误判。所谓稳定，就是在各次测验中都维持同一个合格标准，不能忽高忽低。要做到合格标准准确稳定，教师首先要对合格与不合格学生的知识结构、能力结构差异有非常清楚的了解，并且能够把握住他们在测验试卷上行为反应的差异特征，能编制出准确区分两类学生的试题。其次是测验编制时要严格按照测验计划命题组卷，靠测验计划的稳定性来维护测验合格标准的稳定。

（三）要客观评价自己的命题技术，合理使用各种题型

定向反应型试题答案唯一、评分客观，在测验中多用定向反应型试题可有效提高测验的信度。但是定向反应型试题命题技术要求比较高，没有一定的命题经验和技术，编写定向反应型试题，有时反而会达不到预期的效果。所以学科教师要客观评价自己的命题技术，在经验不足、技术不够的情况下，不必勉强编写定向反应型试题。研究发现，由于教师熟悉教材，熟悉学生，所以只要教师精心编题，精心评阅，教师自编测验中用自由反应型试题施测，同样会有比较好的测验效果。

（四）要注意总结命题经验，提高命题技术

学科教师不是命题专家，但是学科教师却经常自编测验，经常命题出卷，而且都会实际施测，这就给教师提供了许多反馈信息。教师应该充分利用这些反馈信息认真总结自己的命题经验，以便提高自己的命题技术。不能要求学科教师一开始就有很高的命题技术，但在多次命题实践之后，通过不断总结成功、反思失败，教师能逐渐掌握命题的各种技巧。总结命题经验还具体包括不断积聚优秀试题，不断充实个人的“题库”。一次未命好的试题还可以经过修改后再用，不断积聚的结果使得教师拥有一个优质“题库”，使得在以后的命题组卷中，有了一批基本试题，可做到胸有成竹，不至于出大的质量问题。

（五）要尽量控制评分误差，防止简单粗糙

教师自编测验应该防止评分简单粗糙，草率从事。要防止简单地以对错判分，特别是对于自由反应型试题评分应该详细分析学生的作答过程，评价他的思维方式和思维水平。要根据学生实际掌握的知识状况和思维能力综合评分。目前的测验评

分中有两种不可取的方法：一种是教师所命试题为高层次教学目标的测试题，而学生仅根据教材、笔记或一些辅导材料中组织好的答案作一番死记硬背，考试时复述一遍，实际上并无形成自己的思维，而教师也评给一个高分，无形中降低了考试的目标层次；另一种情况是，教师出了一道问答题，而学生不加思维，不加组织，把各种挨得着边和挨不着边的内容一一罗列，教师评分时也不去评价学生的思维，不认定他实际上是知识不扎实，反而到学生漫无目标的作答中选择“正确点”，按点记分，结果还能得高分，甚至得满分。这都是评分简单粗糙的表现。

因此，教师自编测验控制评分误差还要注意两点：一是要在测验前作好参考答案和评分标准，阅卷时要严格按标准评阅。二是要注意防止产生评分误差的名片效应、光环效应、对比效应、先后效应等各种心理效应的影响，还要防止被学生的文字组织和卷面书写风格干扰而远离评分标准的现象。

（六）要做一些定量分析研究

教师自编测验不要求在施测前拿出信度、效度指标，但在施测之后教师对试卷做一些定量分析研究是很有益处的。定量分析内容主要是计算试题的难度、区分度、选择题各选择支的选答率，以及整份试卷的信度，有效标的话还可以计算效度。根据计算数据分析评价各试题质量，评价试卷的整体质量。这是总结命题经验，提高命题技术的更易见效的措施。具体计算分析方法本书有关章节已有介绍，在此不再重复。

练习与思考

1. 结合实际的测验例子试述学绩测验的性质和作用。

2.*就《心理与教育测量》前10章内容，设计一份测验编制计划。

3. 标准化学绩测验的标准化要求有哪些？如何才能做到？

4. 举例说明诊断测验如何实现它的诊断功能？

5. 题库的基本要求是哪些？您认为建设题库有什么好处？

6. 就您的经验谈谈教师自编测验存在哪些不是？改进的途径是什么？

第十二章　能力测验（上）

本章提要：

- 传统智力测验所依据的各种智力理论
- 传统智力测验的评价
- 传统智力测验典型量表的性能及使用：
  比内量表、韦氏量表、瑞文推理测验等

第一节　智力测验的一般问题

一、关于智力实质的理论研究

智力的实质究竟是什么？这是多世纪以来智力研究者们关注的焦点。智力测验虽已有近百年的历史，但对这一问题至今仍是百家争鸣，没有定论。

19世纪末，冯特、高尔顿、卡特尔和比内（Wundt.W、F.Galton、R.B.Cattell、A.Binet）等许多心理学家试图用实验或测验手段评估个体智力。但是，他们当时评估智力时对智力所下的操作定义往往与其对智力的理性理解不相符合，因而，他们的各种智力评估手段的结构效度很低。这种局面迫切要求对智力的实质作出明确界定，从而能够有效地评估智力的个体差异。于是，心理学家们开始致力于关于智力实质的研究。这样的研究几乎贯穿整个20世纪，各种智力理论及其评估手段层出不穷。其中，基于心理测量学、认知科学和生物学的三类研究是最富影响力和竞争力的。

（一）智力的心理测量学理论

1. 智力二因素论

英国心理学家斯皮尔曼（C.Spearman，1904）首先发现一系列心理能力测验之间存在普遍的正相关，并首先利用因素分析方法，将这些相关归因于一种一般因素或共同因素的作

用。他命名这种因素为G因素，并从3个方面对它定义：经验的领会、关系推断和相关推理。他认为G因素对于同一个体是稳定的，它渗透于所有与智力行为有关的任务之中，是一切心智活动的主体，个体间智力的差异就决定于个体拥有的G因素量的多寡。

后来，由于测验间并非完全相关，因而，出于统计上相关分析的需要，斯皮尔曼又提出还存在特殊因素（S因素），并由此构成他的智力二因素论。但他始终强调G因素是智力的核心，而S因素只有在某些特殊情况下（特殊工作或特殊活动）才会表现出来，因而只具有偶然的意义。

2. 智力多因素论

由于统计学中因素分析法的发展，美国心理学家瑟斯顿（L. Thurstone，1938）利用多重因素分析方法首先提出：智力的核心不是单一的G因素，而是许多主要的、基本的、彼此相关的能力因素群。经过多年研究，他和他的学生从56种不同的测验中，分析出语文理解、言语流畅性、推理、空间表象、数字、记忆和知觉速度等7种最主要、最基本的心理能力。

一战以后，吉尔福特（T. Guilford）以20年时间孕育出一个新的智力结构模型，通常被视为瑟斯顿理论的扩展（见图12.1）。在此，智力包括3个维度：

（1）内容因素，指引起心智活动的各类刺激，包括①视觉（F）——形状大小、颜色等；②听觉（A）——声音信号；③符号（S）——字母、数字等；④语义（M）——词句的意义、概念等；⑤行为（B）——各种行为模式。

（2）操作因素，指由各种刺激引起的反应方式与心理过程，包括①认知（C）；②记忆（M）；③发散思维（D）；④聚合思维（N）；⑤评价（E）。

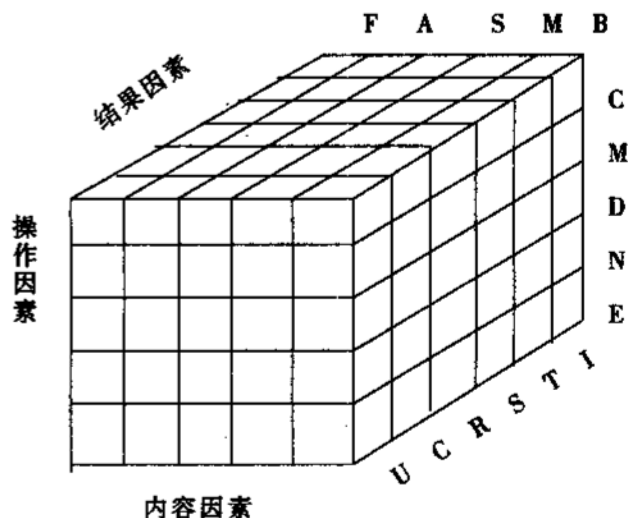


图 12.1　智力三维结构模型

（3）结果因素，指心智活动的产物，亦即对各类刺激的反应结果，包括①单位（U）——可以按单位计算的产物，如一个词、一句话等；②类别（C）——将事物分类；③关系（R）——推断两个事物间关系；④系统（S）——推断一个系统内诸事物的关系；⑤转化（T）——对事物认识的迁移；⑥涵义（I）——解释内涵。吉尔福特认为这三个维度的变化组合可以产生 150 种心理能力。

事实上，智力多因素论者虽然强调智力由多种能力因素构成，但他们后来也不否认 G 因素存在的可能性，只是否认其重要性罢了。

3. 智力层次理论

美国心理学家弗农（P.A.Vernon，1971）提出了智力三层次模型，认为 G 因素处于智力结构最高层，贯穿于其他所

有智力因素之中；第二层是言语能力和操作能力两个因素，分别控制着第三层的主要心理能力，如数学、语文、空间知觉等(见图 12.2)。

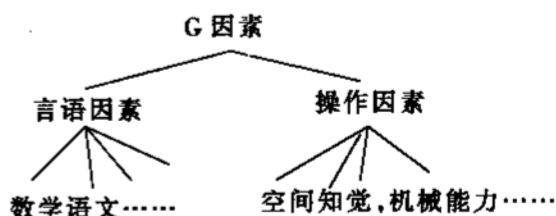


图 12.2　智力三层次模型

(二) 智力的生物学理论

随着神经生理学和生理心理学的发展和成熟，智力的生物学研究在智力领域始终占据一席之地。高尔顿、桑代克(E.L.Thondike)、艾森克（H.Eysenck）和詹森（A.R.Jenson）等人皆从生物学观点出发，认为：智力在人类脑的结构、生物化学、生理学、遗传学等先天因素的影响下形成和发展，它使人类区分于其他生物，同时也使人类个体差异得到反映。詹森的智力振荡理论在其中最具代表性（见图 12.3)。



图 12.3　智力振荡理论的等级双向树图

图中黑点表示大脑皮层的激活点，其数目与智力任务中的

物理刺激数目相对应。图中所标数字表示神经传导链的水平数。

振荡理论假设：每个结点的激活水平是振荡的，因此这些结点有一半时间处于不应期。对结点的刺激若超过了其激活阈限，则将沿着结点链传递下去直至最后的反应通道。因而，对刺激作出反应的时间量，实质上依赖于两个因素：①激活传导所必需经过的链的水平数；②结点的平均振荡周期。个体在这两个因素上的差异，导致了个体的反应时差异，并最终反映了个体在智力上的差异。

可见，詹森强调速度因素在智力上的重要性。事实上，他也承认心理测量学中智力 G 因素的存在，所不同的是他将 G 因素定义成了反应速度。

（三）智力的认知心理学理论

60 年代，认知科学兴起。此后，由于它的影响力和渗透力，越来越多的心理学家开始在信息加工的理论框架下，试图探讨人类智力的内部信息加工机制与过程。在这一领域中具有代表性的研究成果，当属美国心理学家斯坦伯格（R·J·Sternberg，1985）提出的智力认知成分理论。

斯坦伯格认为智力结构由“成分”组成。所谓成分，就是对物体或符号的内部表征进行操作的基本信息加工过程。据成分的概括水平或功能可对其进行不同分类：

1. 据成分概括水平分类

可分为一般成分、类成分和特殊成分。一般成分指所有智力任务操作所必需的成分，类成分指至少两种任务必需的解决某类任务的成分，特殊成分只是单一任务操作所需的成分。斯坦伯格以一个等级结构来说明这三种成分之间的关系，但并未对每类成分的具体内容作进一步的诠释。

2. 据成分功能分类

可分为操作成分、元成分和知识获得成分。操作成分是智力任务完成过程中实际施行的加工过程，其中最普遍存在的信息加工成分有：编码、关系推断、相关推理、应用、比较、证实、反应。元成分是指问题解决过程中使用计划、监控和决策的高级执行过程，其功能包括：审阅问题；选择信息加工成分；选择信息的一种或多种表征；选择信息加工成分的组合策略；决定注意资源的分配；问题解决过程的监控及结果的检验和评价。知识获得成分是指用于获得新知识的过程，包括学习成分、保持成分和迁移成分。三种主要功能成分相互作用，彼此激活或给予反馈（直接或间接），处于一种动态结构之中。

二、关于智力评估的实践探索

受达尔文进化论思想的影响，英国心理学家高尔顿将智力归诸于遗传的素质，成为智力的个别差异研究和科学测量智力的主要创始人。1884年，高尔顿开始运用实验手段测量智力，结果以反应时表示，并且首先发现反应时与教师评定的智力等级间的正相关。在智力的早期研究中，卡特尔、桑代克等人皆沿袭了高尔顿的实验室方式，认为反应时与其他简单的感知觉辨别测验相结合可以评估智力的个体差异。

1904年，斯皮尔曼提出了智力G因素的存在。这一理论观点不仅对以后关于智力实质的理论研究产生了巨大的影响，而且也成为智力测验产生的理论基础。1905年，法国心理学家比内和医生西蒙（T.Simon）在智力G因素论的影响下，合作制成世界上第一个智力量表，以测验的总分或平均分作为个

体智力G因素水平的评估指标，并以此标定智力的个体差异。从此，比内—西蒙智力量表便作为智力测验的传统模式而存在。在随后的50～60年时间里，行为主义学派在心理学中占据着主导地位，心理学家们的研究兴趣更多集中于行为的结果而非其内部过程。因而，这段时期内的智力评估，几乎皆以智力的心理测量学理论为基础，并遵循着比内—西蒙量表的传统——只是测验内容的细节不同，并且评估指标几经改进之后，离差智商成为最广泛使用的指标。

智力测验一产生，便被迅速地应用于人类社会的各个方面，并且，作为一种度量工具，它们在对个体的分类和预测上显示了非凡的使用价值。然而，智力测验同时也遭到来自各方面的批判。其中，最激烈的批判之一是认为智力测验过分注重于个体的知识结构，而知识是教育的结果，教育又极具特定社会和文化背景的影响，因此，测验的应用受到歪曲或限制。批判者们提倡文化公平测验，这种思想集中体现在智力的生物学研究之中。

关于智力的稳定的遗传力的研究、智力与大量生物学指标之间的相关研究，以及智力与反应时之间的相关研究为詹森的理论假设提供了实验证据，并表明了速度对于智力G因素的重要性。因此，詹森认为可以设想从更为简单的信息加工现象入手，以一系列不受社会和文化背景影响的纯粹的反应时测验来取代传统的智力测验，并以个体在测验中的反应时指标来标定其智力G因素水平。

由于传统的智力测验和詹森所倡议的反应时测验均注重于个体在测验中的行为结果，并以一个总的指标来评估智力的个体差异，而未在更为精细、严密的水平上对个体心理活动过程的内部加工机制进行分析，因此，60年代以后的认知心理学家们对此提出异议，并开始寻找新的智力评估方法。

在认知心理学中的减法反应时法的启发下，斯坦伯格提出了智力的认知成份分析法。这种方法从复杂认知作业的操作入手，并在理论上假设任何一个复杂的问题解决都由一系列基本的认知操作成分（如编码、推断、应用等）构成，然后通过精心设计的反应时实验，分解出不同智力水平的个体解决同一复杂问题所采用的各种认知加工成分，并记录每一个体在每一加工成分上的反应时参数，最后比较个体和总体的各成分参数，就可以评估个体在每一加工成分上的水平高低，从而能够相当精确地诊断出个体认知过程中真正的、内在的薄弱点，并因此而能对症下药，及时有效地给以补救和引导。

然而，以反应时表示的信息加工速度虽然可以有效地反映个体操作成分上的差异，但棘手的是：人们可以主动地有意识地控制加工速度，进行合理的资源分配，平衡速度和准确性的矛盾。这正是斯坦伯格所说的智力元成分的功能。显然，信息加工速度并非元成分的主要特征，因而以反应时作为元成分的评估指标是无效的或至少是不足够的。那么应该如何评估元成分？这个问题目前仍在研究和探索之中。

事实上，以詹森为代表的生物学智力论者和以斯坦伯格为代表的认知心理学智力论者虽然从不同的角度指出了传统智力测验的不足，并在各自的理论基础上对智力评估提出了新思路、新方法，但是这些新思路和新方法往往还只停留在设想或实验研究的水平之上，而未能制订出现成的、切实可信的智力测量工具，因而便无法被应用于实际之中以满足社会的需要。到目前为止，在社会各界用以评估智力个体差异的测量工具中，影响最大、普及面最广、权威性最强的仍是传统的智力测验。

三、传统智力测验的若干问题

(一) 传统智力测验的结构效度

由于传统的智力测验是在智力 G 因素论的基础上编制，因此，若测验具有较好的结构效度，那么我们便认为它基本测出了个体在智力 G 因素上的水平高低。然而，如果我们作进一步的讨论："智力 G 因素真正存在吗?""智力 G 因素的实质究竟是什么?""智力 G 因素上的个体差异意味着智力全部的个体差异吗?"那么便会遗憾地发现对于这些问题的讨论始于很多年以前而至今仍无定论。若想解决这些问题，唯一有效的途径就是统一对于智力实质的看法。这就意味着我们前面所述的各种不同的智力研究方向将向一个共同点汇合。

分析智力理论的研究趋势，各种智力理论研究方向之间的结合是可能的并且势在必行。早在 1957 年，美国心理学家克伦巴赫（Cronbach）就提出：科学心理学应当将相关研究和实验研究有机地协调起来。以心理测量学为代表的相关研究能够揭示智力任务上各种不同智力因素间的关系，但它却不能解释各种智力因素及其相互关系的内在加工实质。以认知心理学为代表的实验研究能够以相对确定的方式揭示认知活动的内在机制，但单从实验本身却难以说明加工系统的各种成分对于完成智力任务是否是一般有效的。因而，完整、充分地认识人的智力需要两种研究相结合。而智力的生物学研究，从辩证法的角度来看，它所强调的智力的遗传素质应该作为智力研究的生理基础而存在。

智力研究者们期望通过共同的努力而最终达到对于智力实

质的共识。那么，关于智力实质的一个综合的统一的理论究竟能不能产生？若能，则什么时候可以产生？这些目前来说均是未知数。因而，关于智力测验的结构效度，目前我们只能做狭义的解释，即以不同智力理论模型为基础的智力测验，其结构效度应具备不同的含义，并且我们最终对测验结果的解释亦相应不同。因此，在选择和使用智力测验时，这是一个必须慎重考虑的问题。

（二）传统智力测验的功能

测验的基本功能是测量个体差异，因而传统智力测验的功能便是对智力的个体差异的测量。大量的测量结果表明：在一般人群中智力高者或低者均占少数，智力中等或接近中等者约占全部人口的 80%，基本上呈常态分布。

由于智力测验所依据的理论上的特性，以及智力分布的常态性，智力测验在实际中常常行使将人群分类的功能。韦克斯勒（D. Wechsler）曾按智商的高低，把智力分成 7 类，如表 12.1。

表 12.1　韦克斯勒对智力的分类

IQ	类别
130 以上	极优秀（天才）
120 – 129	优秀
110 – 119	中上（聪颖）
90 – 109	中材
80 – 89	中下（迟钝）
70 – 79	低能边缘
70 以下	智力缺陷

此外，由于传统智力测验结果与学习成绩、教师评定等外在效标间的高的正相关的存在，智力测验在实际中又常被作为

预测源测验行使预测的功能。

社会对智力测验的需求不仅仅是将人群分类并预测个体未来可能的成功程度，还要求测验能进一步对个体进行诊断和采取相应且及时的补救措施。而智力的心理测量学概念强调G因素的存在，只从宏观上描述个体的外在行为结果，却忽视了对个体心理活动过程进行精细的分析及对内部加工机制的探讨，因此，以此为依据的传统智力测验可以对人群进行分类和预测，却无法对个体真正的内在的薄弱点作出精确诊断，从而不能满足社会更为精细的需求。

（三）传统智力测验的公平性

利用测验评估智力差异时，首要前提是必须客观公正。传统智力测验对任何人都是公正的么？这是一个长期以来颇有争议的问题。对此持否定态度者主要从以下几个方面提出质疑：

1．性别差异

对于智力的性别差异的研究，已有多年的历史。研究者们通常会发现男女两性在智力上的差别在统计上并不显著，因此，就整体而言，智力很可能并不存在性别差异。然而，具体到对智力的不同能力因素上的研究，则结果一般表明存在性别差异：男性在数学推理、视觉——空间能力、躯体运动速度和协调方面优于女性；而女性在言语流畅性、言语理解和记忆等方面则优于男性。这样一来，智力测验本身的内容和结构，对男女两性便未必公平了。一般测验都包括多种性质的题目，有些可能更适于男性，有些则更适于女性。如果测验中所包含的利于男性和利于女性的题目并不均衡，那么此测验对其中某一性别的人群便是不公正的，此时便应对测验结果所显示的性别差异作出慎重解释。传统智力测验中的《韦氏成人量表》的题目就存在这样的不平衡问题。

2. 职业差异

关于智力的职业差异的一类研究表明：不同职业的人在智商上具有差异，从事专业工作的人员（如会计师、律师、工程师）的智商最高，而诸如工人、农民这样的劳动者相对则智商最低。有些资产阶级学者据此得出“劳动人民天生愚笨”的结论，为其阶级剥削提供依据。显然，这是一种阶级偏见。在当代社会，职业分工日益精细，不同的职业需要不同的能力是一个公认的事实。而传统智力测验并不能全面反映人的各方面能力，因此仅以智商高低来判断人的智愚并不充分，由此得出“天生如何”的结论更是荒谬。智力受到遗传与环境的双重影响，劳动人民及其子女之智商若相对偏低，则很可能是由环境影响所致，而不能断言其天生如此。

3. 文化和教育差异

传统智力测验经常受到的批评是：个体在测验上的得分往往受知识经验的影响，因此测验对不同文化背景或教育水平的团体是不公平的。大量测验结果显示了显著的城乡智力差异和种族智力差异，后者曾为种族歧视者所利用，成为其种族歧视的借口。

然而，智力测验结果反映的城乡差异和种族差异，并不能完全归因于城乡或种族本身由遗传引起的差异。事实上，文化和教育因素在此很可能起了主导作用：一方面是由于构成智力测验的题目本身在很大程度上是对个体受文化影响和受教育程度的测定；另一方面是由于城乡之间或不同种族之间的生存环境在文化和教育方面有极大的差异，城市儿童或白人儿童一般生活在浓厚的文化氛围之中，并且其家庭的社会经济地位普遍足以为其提供较好的教育，农村儿童或黑人儿童则不然。此外，另一些研究事实亦为此提供了依据：一是现代传播媒介以及各种信息交流手段的日益丰富，使得美国农村儿童对文化的

接触日益广泛，智力的城乡差异在明显缩小；二是一些研究者通过人为改变某些黑人儿童的生活环境，给予他们较好的教育和一定的文化熏陶，一段时间以后发现他们的智力水平明显上升。

为了保证智力测验对不同文化背景团体的公正性，很多人试图编制排除文化影响的“超文化”测验或所包含的文化因素适宜于不同团体的“文化公平”测验。现在关于这方面的研究取得了一定效果，但目前为止，还没有出现一个成功测验可以用来取代现有的传统的智力测验。

（四）传统智力测验的预测效度

在用智力测验对个体未来可能成功程度作预测时，一般都假定所测的智力是个体相当稳定的特质。而事实上，人的智力并非一成不变，它会由于某些因素的影响而发生变化。智力的可变性主要表现于以下几个方面：

1. 智力随年龄成熟而发展

一般研究表明：人类智力随年龄而增长。绘制智力与年龄关系的曲线。可以看到：智力在童年期急速增长，在青少年期增长稍缓，在成年期达最大值，保持稳定一段时期后开始有所下降。（见图 12.4）

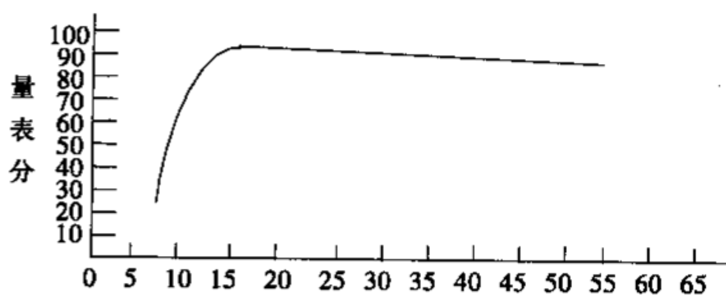


图 12.4　智力成长曲线（根据韦克斯勒 1958 年的研究）

虽然对于智力成长曲线的研究结果尚不尽一致，但大概可以归纳如下：

(1) 智力发展在十二三岁以前呈直线上升趋势，十三岁后开始减慢。

(2) 个体智力水平与其智力发展速率及停止年龄密切相关：一般智力高者发展速度快而停止年龄晚；智力低者则发展速度慢且停止年龄早。

(3) 智力发展高峰期虽存在个体差异，但总的来说，早期研究认为一般人的智力在16～18岁时到达顶峰，近期研究结果又将之推迟至25岁。

2. 智力随环境而变化

在遗传素质确定的情况下，环境的变动将对个体智力产生一定影响。比方说，突然从经济条件良好且教育环境良好的状态下，转到贫穷和无法接受良好教育的境况，或从文化和物质均较贫瘠的农村某地迁人文化氛围浓、教育水平高的城市中生活，都会对个体智力产生消极或积极的影响。

3. 智力随个性特质的不同而产生不同变化

个体人格特征也是影响智力发展的一个因素。例如，韩(Haan, 1963) 的研究结果表明：智商的改变与个人的心理防卫机制有密切关系，凡是运用退缩、否认、合理化的人，其智商有降低的趋势；反之，运用客观、建设性、面对现实的人，其智商有上升的趋势。

由于上述智力的可变性，利用智力测验来预测个体今后的可能成功程度的效能便有被夸大的可能。一个在当前测验上IQ较低的个体，并不一定今后就笨而且没有成就。比如，一般来说女性无论在生理上还是智力上发育均较男性为早，那么，女性早期在智力测验上表现出的优越性、并不能说明女性

今后一定比男性成功。因此，以智力测验预测个体未来成就时，一定要慎重考虑其预测效度问题。

虽然传统智力测验在理论基础和实际运用中存在这样或那样的问题，以致招到来自各方的批判，甚至一度受到社会的抵制，但是，到目前为止，智力测验仍被作为有效的智力个体差异评估工具而得到广泛应用。究其原因，大约可以归于以下几个方面：

首先，智力本身虽具有可变性，但从另一个角度来说，它也具有稳定性。其稳定性主要表现在：个体智力在其相应团体中的相对位置长时期内保持稳定。这种相对地位的稳定可以首先归因于个体的遗传素质。研究表明，血缘关系越近的人智商相关越高，可见遗传对于智力的发展具有不可忽视的作用。虽然环境的变化对智力发展会产生很大的影响，但就普遍范围而言，多数人的环境是相对稳定的，突如其来的环境巨变相对少见。而且，后天经验是一个积累的过程，先前经验为以后的发展提供了基础，因而最初发展较快、智力水平较高的个体很有可能在其团体中继续保持领先地位。

可见，个体智力的相对稳定性为智力测验具有一定预测效度提供了可能性。并且，由于个体智力的发展到一定年龄以后会越来越慢，最终会达到顶峰并在此后长时期内处于稳定状态，所以智力测验的预测效度便会出现随受测者年龄递增的趋势。

其次，智力测验实际运用于选拔和安置人员时，往往被实践证明其对学生和职业等效标的预测具有较好的效度，因而可以有效地帮助决策者提高决策正确率。

再次，由于智力测验对不同团体可能存在的不公平性，人们已经试图从改善智力测验本身来缓解这一问题，比如改善题目结构，或据亚文化群的特点为不同团体编制不同的测验，或

在同一测验中为不同团体制定子常模等。但是，事实上，从另一角度来说，当智力测验被用于选拔人员时，我们更应看重的是其预测效度而非其公平性。只要一个测验确实能够在一定的录取率下相当准确地筛选出最有可能成功的人，那么该测验就应是可行的。至于它对各种不同团体公平与否的问题，最根本的解决办法还是建立一个政治、文化、经济等各方面高度平等的社会，从本质上消除文化、经济、教育方面对某些特殊团体的不公平。

最后，虽然人们指出传统智力测验的种种不是和局限，并且从各种角度提出了更全面更完善的智力评估手段的设想，也有很多人在实践中作了诸多尝试，但至今仍未出现成熟的、超越于传统智力测验之上的智力评估工具。因此，传统智力测验在智力评估中的地位目前仍是不可取代的。

第二节 个体智力测验

传统智力测验由于施测对象的不同可以分为个体智力测验和团体智力测验，前者一般由一位主试对一位被试进行面对面的施测，后者则可由一位主试同时对若干被试进行施测。本节将对代表性最强、影响最大的个体智力测验作一简要介绍。

一、比内量表

（一）比内—西蒙量表

1.1905年量表

这是比内和西蒙出于诊断异常儿童智力的需要，于1905年编制而成的世界上第一个智力量表。它包括30道测验项目，种类繁多，可以测量智力的多方面表现，比如记忆、言语、理解、手工操作等。它以通过多少项目作为区分智力的标准，并且显现出年龄量表的雏型，比内和西蒙在此已指明不同年龄的儿童所能通过的项目。

2.1908年量表

这是第一个年龄量表。比内和西蒙在此对1905年量表作了如下修订：①测验项目增至59个；②测验项目以年龄分组（3～13岁，每岁一组）；③以智力年龄来评估个体智力，即儿童最后能通过哪个年龄组的项目，便说明他具有这一年龄的智力水平，而不论他的实际年龄是多少。

3.1911年量表

比内在1908年量表基础之上对其做最后一次修订，除了改变一些项目内容及其顺序之外，还将其适用范围扩大，增设了一个成人题目组。

虽然如今比西量表由于其简陋和非标准化而不再为当代人所使用，但它在智力测验历史上的贡献是不可磨灭的，它的主导思想成为其后智力测验所遵循的传统。

（二）斯坦福—比内量表

1. 斯坦福—比内量表的发展

比西量表发表以后，戈达德（H. Goddard，1908）第一个将其介绍到美国。此后，又有一些人对它进行了修订，其中美国斯坦福大学的推孟（L. Terman）教授的工作最负盛名。

（1）1916 年量表。

推孟 1916 年发表的斯坦福—比内量表（简称斯—比量表）中，对于比西量表中的项目或者保留，或者修改，或者删除，并在此基础上又增设了 39 个新项目。该量表首次引入比率智商的概念，开始以 IQ 作为个体智力水平的指标。而且，为了使测验标准化，该量表对每个项目施测规定了详细的指导语和记分标准。

（2）1937 年量表。

推孟 1937 年对斯—比量表作了第一次修订，修订后的斯—比量表由 L 型和 M 型两个等值量表构成。该量表适用年龄由 1916 年的 3～13 岁扩展到 1.5～18 岁，并在修订时选取了更大的代表性样本以获得其信度、效度资料，不过其样本仍局限于白人，且偏重于社会经济地位较高家庭的儿童，因而仍未能全面反映美国当时人口状况。

（3）1960 年量表。

这个量表汇集了 1937 年量表的 L 型和 M 型中最佳项目而成 LM 型单一量表，适用于 2 岁到成人。该量表的重大改变在于舍弃了比率智商，引入了离差智商概念，以平均数为 100，标准差为 16 的离差智商作为智力评估指标。

（4）1972 年量表。

此量表保持 1960 年量表的测验内容不变，重新修订常模，所选常模团体包括了美国各地区、各社会阶层、各种经济状

况、各民族的2100名儿童，取样代表性有了很大提高。

2. 斯比量表的信度与效度

（1）信度。

一般说来，斯比量表对年龄大的被试比年龄小的被试信度高，对于智商低的被试比智商高的被试信度高。计算其L型和M型量表的复本信度，在2.5~5.5岁为0.83~0.91，在6~13岁为0.91~0.97，在14~18岁为0.95~0.98（下限信度值来自于IQ为140~149的被试，上限信度值来自于IQ为60~69的被试）。再测信度与复本信度的研究结果大体一致。

因此，总的看来，斯比量表是一个高信度的测验，各种年龄和IQ水平的信度系数大都在0.90以上，意味着在被试实得分数变异中90%以上来自于真分数变异，而由随机误差引起的分数变异不足10%。

（2）效度。

斯比量表的效度具有多方面证据：

内容效度：斯比量表中所包含的项目涉及到多方面的内容，如言语、类比推理、理解、记忆、空间关系、数字等，而这些内容又被公认在智力范畴之内。

效标关联效度：由斯比量表而得的智商分数与学业成绩、教师评定、受教育年限等外在效标分数间存在普遍正相关，效标关联效度系数大多介于0.4~0.75之间。由于斯比量表以文字材料为主，因此它对言语方面的预测有效性较之其他方面更高一些。

结构效度：斯比量表的理论构想主要基于以下两方面：①智力随年龄而发展，其成长曲线特征为先快后慢；②智力结构中存在一般因素G，它渗透于每一智力行为之中，是智力的核心。斯比量表对于其理论构想的测量有效性已得到一定程度的证明：一方面，斯比量表的信度研究显示出其再测稳定性程度

随年龄而提高的趋势，从而表明智力随年龄而先快后慢发展的特点；另一方面，在1960年量表中，虽然每一项目涉及不同智力行为，但项目分析结果显示各项目与测验总分的平均相关系数为0.66，这表明各项目所测的特质同质性很高，因而正是支持了其理论假设中贯穿于所有智力行为之中的智力G因素的存在。

（三）中国比内测验

从本世纪20年代起，我国心理学家陆志伟和吴天敏便开始斯坦福—比内量表的中国版修订工作。1924年，陆志伟在1916年斯比量表的基础上修订而成《中国比内西蒙智力测验》。1936年他和吴天敏合作发表第二次修订本。1978年，吴天敏主持第三次修订，1982年完成《中国比内测验》。

该测验共有51道题，从易到难排列，每题代表4个月的心理年龄，这样从2～18岁，每个年龄段有3道题。不过最后的智力评定指标并非智龄，而是离差智商。

施测时，先根据被试年龄从测验手册的附表一中查到开始作答的题号，如2～5岁儿童从第一题开始作答，6～7岁儿童从第7题开始作答，8岁儿童从第10题开始作答，等等。然后根据指导语进行逐题测试，采用全或无的记分方法，即通过为1分，不通过为0分，连续5题得0分便停止测验。最后根据测验总分和被试实足年龄，可从指导手册的常模表中查得被试的智商，如某4岁零3个月的儿童得分为9，则可知其智商为108。

中国比内测验必须个别施测，并且要求主试必须受过专门训练，对量表相当熟悉且有一定经验，能够严格按照测验手册中的指导语进行施测。

为了节省测验时间，吴天敏在《中国比内测验》的基础上

又制定了一份《中国比内测验简编》，由8个项目组成，通常只需20分钟即可测完。

二、韦克斯勒量表

（一）韦氏成人智力量表

1．韦氏成人智力量表英文版

（1）韦氏成人智力量表的产生与发展。

●韦克斯勒——贝尔韦量表

美国心理学家韦克斯勒在临床心理工作中发觉斯比量表在成人智力水平评估上的不足，他认为斯比量表的内容和题目是针对儿童设置的，过份强调速度而又缺乏难度，对成人而言表面效度很低，无法引起成人的兴趣，而且斯比量表的常模资料亦来自儿童，智龄的概念也不适用于成人。因此，他于1934年开始致力于智力测验的编制和研究工作，1939年发表了韦克斯勒——贝尔韦智力量表Ⅰ型（Wechsler—Bellevue Scale Form I，W－BⅠ）。

W－BⅠ是第一个成人智力测验，它的内容是以特别适合成年人使用的眼光来选择的，并用一系列不同的子测验的形式来编制整个测验，每个子测验内的题目皆由易到难顺序排列。

由于W－BⅠ在常模样本的代表性及子测验信度上的不足，韦克斯勒又于1949年增加了Ⅱ型（W－BⅡ）。W－BⅠ和W－BⅡ主要用于测量10～60岁的被试，它们在内容和形式上为后来发展的各种量表奠定了基础。

●韦氏成人智力量表修订版

韦克斯勒对W－B做了修订和重新标准化，于1955年编

制出版韦氏成人智力量表（WAIS），1981 年又出版了再次修订和标准化后的 WAIS，称为韦氏成人智力量表修订版（WAIS－R）。

WAIS－R 和 W－B 及 WAIS 一样由 11 个分测验组成，其中常识、背数、词汇、算术、理解、类同等 6 个分测验又构成言语分量表，填图、图画排列、积木图案、拼图、数字符号等 5 个分测验构成操作分量表。在此，每个分测验内的题目由易至难排列。并且，言语测验和操作测验交替施测。

WAIS－R 的每个分测验独立记分，再转化为平均数为 10，标准差为 3 的标准分数。六个言语分测验的标准分数相加可得言语量表分，五个操作分测验的标准分数相加可得操作量表分，所有分测验的标准分数相加可得全量表总分。最后，将这些量表分数转换成平均数为 100，标准差为 15 的离差智商分数，便可得到言语智商、操作智商和总智商。

WAIS－R 的常模团体由 1880 人组成，男女各半，分配在 16～17，18～19，20～24，25～34，35～44，45～54，55～64，65～69，70～74 岁 9 个年龄组。韦克斯勒非常注重取样代表性，尽量使之与美国 1970 年人口统计资料中的各种比例相符。他根据常模团体的测验结果，为每个年龄组分别制定常模。因此，根据被试的原始分数查得的言语、操作和总的智商分数，表明了被试在他所属的年龄组团体中所占的相对位置。

（2）韦氏成人智力量表的信度和效度。

●信度

WAIS－R 的手册中报告了 11 个分测验以及言语分量表、操作分量表和全量表在各个年龄组上的信度资料。其中背数和数字符号两个分测验计算的是复本信度，其余均计算分半信度。结果表明：全量表的信度在各年龄组上的分布为 0.96～0.98，言语量表的信度分布为 0.95～0.97、操作量表的信度

分布为0.88～0.94。分测验的信度相对低一些，但11个分测验在各年龄组上的89个信度系数中也只有5个低于0.70，最高也达到0.96。

●效度

WAIS－R没有收集效度资料，但韦克斯勒等人曾对WAIS的效度作了大量研究。

结构效度：韦克斯勒曾明确指出："WAIS中的11个分测验是从各个方面来测量智力，而不是测量不同类型的智力。"他认为，"智力是个人有目的地行动，理智地思考以及有效地应付环境的整体的或综合的能力。"对WAIS的因素分析结果表明：在测验分数的总变异中，有50%的变异来自智力一般因素。在WAIS－R中，各分测验之间和分量表之间存在的普遍的、显著的正相关，也表明智力G因素渗透于智力行为的各个方面。

内容效度：韦克斯勒在量表中设计的11个分测验，均取自于前人，它们在早期智力量表中皆被成功地使用过，并在临床实践中显示了它们的价值，被公认为智力行为的主要范畴。

效标关联效度：在异质性较高的团体中，WAIS与斯比量表的相关在0.80左右。WAIS与各种教育与职业效标间也有相关，如文职人员的言语智商平均高于其操作智商，技术工人则正相反。

2. 韦氏成人智力量表中国修订本

1982年，在湖南医学院龚耀先主持之下修订出版了WAIS的中国修订本（简称WAIS－RC）。

（1）WAIS－RC的修订工作。

WAIS－RC在项目内容上变化不大，只是删除了部分完全不适合我国文化背景的题目，并根据我国常模团体的测验结果对测验项目顺序作了适当调整。其主要内容如下：

●言语量表

常识测验：共29题，内容取样范围极广，尽量避免涉及专业领域的内容。例如："钟表有什么用?""我国首都在哪儿?"等，结果以"1""0"计分，用于测量被试的一般智力因素和记忆能力。

理解测验：共14题，要求被试说明在某种特定情形下应做什么，或解释一些话的意思。如："为什么不要同坏人交朋友?"等，以"0""1""2"方式计分，用于测量被试运用实际知识解决问题的能力和社会适应能力。

算术测验：共14题，内容属小学算术范围，如"8个人在6天内可以完成的工作，若半天内必须完成，应找多少人来做?"题目限时完成，以"1""0"计分，用于测量被试基本数理知识和数学推理能力。

类同测验：共13题，要求被试说出两件事或物的相似之处，如"斧头—锯子"，依被试回答的全面程度以"0""1""2"计分，用于测量被试抽象逻辑思维和分析概括能力。

背数测验：由主试口述一串由3~12个数字随机排列组成的数字系列，要求被试按顺序复述，共12题；再由主试口述一串由2~9个数字随机排列组成的数字系列，要求被试倒着复述，共10题。结果以"0""1""2"计分，用于测量被试的注意力和短时记忆能力。

词汇测验：主试将一张包括40个词汇的词表呈现在被试面前，要求被试指出主试所读的词，并对其意义进行解释。结果以"0""1""2"计分。用于测量被试的言语理解能力。

●操作量表

数字符号测验：呈现数字与符号的对应样例：1~9每个数字对应一种符号。要求被试根据样例在每个数字下填上相应的符号，限时进行，以"0"、"1"计分，用于测量被试建立新

概念的能力和知觉辨别速度。

填图测验：共21张画片，每张图上都有缺失的部分，例如：人没有耳朵，动物没有尾巴等，要求被试指出缺失的部分，以“0”“1”计分，用于测量被试视觉记忆与辨别能力。

积木图案测验：给被试9块积木，每块各面分别涂有全红、全白或半红半白的颜色；同时给被试呈现10个图形，要求被试在限定时间内用积木拼摆出所呈现的图形。此分测验主要用于测量被试视知觉组织、视动协调及分析综合能力。

图片排列测验：共8组图片，每组图片打乱顺序后呈现给被试，要求被试重新以适当顺序排列，以组成一个连贯的故事情节。用于测量被试分析综合和知觉组织能力。

拼图测验：要求被试将一个被切割成几块的图形拼好，根据被试完成的速度来计分，用于测量知觉组织及概括思维能力。

WAIS－RC建立了农村和城市两个常模，从16岁至65岁以上共分8个年龄组，人口组成情况主要依据长沙市及其郊区的有关资料，不过实际取样来自21个省。

（2）WAIS－RC的信度和效度。

对WAIS－RC的信度研究表明：各分测验的分半信度在不同年龄组的分布为0.30～0.85之间，各分量表和全量表的再测信度在0.82～0.89之间。

对WAIS－RC的效度研究表明：在高考成绩上差异显著的被试，在WAIS－RC测得的智商上同样表现出显著差异，说明WAIS－RC具有一定的效度。

（二）韦氏儿童智力量表

1．韦氏儿童智力量表英文版

（1）韦氏儿童智力量表的产生与发展。

韦氏儿童智力量表（WISC）是韦氏成人智力量表向较低年龄水平的扩展。它是1949年由韦克斯勒在W-BⅠ的基础上修订而成。它基本上保留了原来的测验形式，只是降低了测验难度，并且增添了一个迷津分测验，用于测量知觉的速度和准确性。它的主要特色在于放弃智龄概念，采用离差智商代替比率智商，并使得离差智商从此成为智力测验中最广泛使用的指标。

韦克斯勒于1974年完成对WISC的修订和重新标准化的工作，发表了韦氏儿童智力量表修订版（WISC-R）。

WISC-R共包括12个分测验，分别构成言语量表和操作量表，其中背数和迷津两个分测验是备用测验，可作为某一同类测验的替换或补充测验。

WISC-R适用于6~16岁的儿童，从6岁0个月到16岁11个月，每四个月为一个年龄组，分别建立了常模表，可直接由原始分查得言语智商、操作智商和总智商。

(2) 韦氏儿童智力量表的信度和效度。

●信度

研究表明：WISC-R中各分测验的分半信度分布在0.70~0.86之间，再测信度在0.65~0.88之间；各分量表和全量表的分半信度在0.90~0.96之间，再测信度在0.90~0.95之间。

●效度

WISC-R的效度证据来自以下几个方面：

效标关联效度：以年龄为效标，可证实WISC-R上的原始分数确实随年龄增长而提高；以学绩测验或其他学业成就为效标，发现WISC-R与这些效标间的相关系数在0.50~0.60之间；以斯比量表为效标，发现WISC-R的总智商，言语智商及操作智商与斯比量表的智商之间在各年龄组的平均相关为

0.60～0.71之间。

结构效度：对WISC－R的因素分析结果和对WAIS的分析极为相似，同样发现了智力一般因素的存在。同时WISC－R中的言语量表和操作量表在各年龄组的平均相关为0.60～0.73，说明二者之间存在许多共同变异，这为智力G因素的存在进一步提供了证据。

2. 韦氏儿童智力量表中国修订本

WISC－R的中译本于1979年由林传鼎、张厚粲等人提出并于1981年底初步完成修订工作。这次修订的重点在于删改一些文字内容和图像，使题目尽可能地适合中国儿童特点，并在此基础上编制中国常模。该测验的常模团体取样来自大、中城市，因而只适用于中等以上城市的儿童。其信度和效度也已在一定程度上得到某些研究结果的支持。

（三）韦氏幼儿智力量表

韦氏幼儿智力量表（WPPSI）出版于1967年，适用于4～6岁半的儿童。

WPPSI同样包括11个分测验，其中3个分测验（句子复述、动物房、几何图案）是为了适应幼儿特点而新编的，另外8个（常识、理解、词汇、算术、类同、填图、迷津、积木图案）则与WISC相同。

WPPSI亦给出言语智商、操作智商和总智商。其常模团体取自美国不同地区、种族和家庭的儿童，每半岁为一年龄组，每一年龄组都建立了常模表。

WPPSI在手册中报告言语量表、操作量表和全量表的分半信度在0.84～0.94之间，再测信度在0.86～0.92之间。对WPPSI的因素分析发现了智力G因素的存在；同时，对98名5～6岁儿童的施测结果表明：WPPSI的各分量表及全量表的

智商与斯比量表的智商的相关系数在 0.56～0.76 之间。上述这些结果为 WPPSI 的信度和效度提供了支持。

韦氏的 3 种智力量表互相衔接，适用的年龄范围可从幼儿直到老年，成为智力评估中最广泛使用的工具。

第三节　团体智力测验

第一次世界大战期间，美国心理学会主席耶克斯（M.R.Yerkes）和桑代克、推孟等许多著名心理学家提出用测验招募和选拔士兵。但面对短时间内动员数百万兵员的任务，采用个别施测的智力测验显然无法完成任务。于是，在推孟的学生奥蒂斯（A.S.OTis）编制的团体智力测验的基础之上产生了陆军甲种测验，后来又针对不识英文或有阅读障碍的人编制出陆军乙种测验。从 1917 年 9 月到 1919 年 1 月，受测者总人数达 170 多万人。陆军测验的成功，使团体智力测验的研究、编制及应用迅速发展起来。本书将简单介绍其中一些影响较大、应用较广泛的团体智力测验。

一、陆军测验

陆军甲种测验由 8 个分测验组成，包括指使测验（照令行事测验）、算术测验、常识测验、异同测验（区别同义词和反义词）、字句重组测验、填数测验、类比推理测验和理解测验

陆军乙种测验属于非文字测验，由7个分测验组成，包括迷津、立方体分析、补足数列、译码、数字校对、图画补缺和几何形分析。

陆军甲种测验的效度资料来自它与军官评定的相关（0.50～0.70），与斯比量表的相关（0.80～0.90），与教师评定的相关（0.67～0.82），以及与学业成绩的相关（0.50～0.60）。陆军乙种测验与甲种测验的相关达到0.80。

二、瑞文推理测验

（一）瑞文推理测验的产生与发展

瑞文推理测验是由英国心理学家瑞文（C.Raven）编制的一种团体智力测验，又称瑞文渐进图阵。它是非文字型的图形测验，分为三个水平：

1. 瑞文标准推理测验

瑞文1938年编制出版该测验，它适用于5.5岁以上智力发展正常的人，属于中等水平的瑞文推理测验。

2. 瑞文彩图推理测验

由瑞文1947年编制而成，适用于幼儿和智力低于平均水平的人，属于瑞文推理测验的3个水平中最低水平的测验。

3. 瑞文高级推理测验

最初编于1941年，经1947年、1962年两次修订成为现在的形式，适用于智力高于平均水平的人，是最高水平的瑞文推理测验。

以上3种水平的瑞文推理测验均由两种题目形式组成，一种是从一个完整图形中挖掉一块。另一种是在一个图形矩阵中

缺少一个图形，要求被试从提供的几个备选答案中，选择出一个能够完成图形或符合一定结构排列规律的图案。

瑞文推理测验的理论假设源于斯皮尔曼的智力一般因素理论。瑞文将智力G因素划分为两种相互独立的能力，一种称再生性能力，表明个体经过教育之后达到的水平；一种称推断性能力，表明个体不受教育影响的理性判断能力。瑞文认为，词汇测验是对再生性能力的最有效测量，而非言语的图形推理测验则是对推断性能力的最佳测量，这就是瑞文推理测验的来源。

瑞文测验的优点在于测验对象不受文化、种族与语言等条件的限制，适用的年龄范围也很宽，从5岁半直至老年，而且不排除一些生理缺陷者。测验既可个别进行，也可团体实施，使用方便，省时省力，结果以百分等级常模解释，直观易懂，因而，该测验在世界各国广泛通用。

（二）瑞文标准推理测验中国修订本

1985年，我国张厚粲教授开始主持瑞文标准推理测验中国城市版的修订工作。

这次修订工作基本保留了原测验的项目形式及指导语。测验共由60道题目组成，分为A、B、C、D、E 5个系列，每一系列包含12个题目。项目系列由易至难排列，每一系列内部的项目亦由易至难排列。每一项目均为“1”“0”计分，最后根据总分查得常模表中相应年龄组的百分等级。

测验常模团体根据人口普查资料取自全国大、中、小城市，取样时注意到性别、文化、职业等人口比例分配，从5岁半到16岁半每半岁为一年龄组，20岁以上每10岁为一年龄组，17岁至19岁为一年龄组，70岁以上为一年龄组。

瑞文标准推理测验中国城市版的分半信度为0.95，再测

信度在0.79～0.82之间。它与WISC－R的中国修订本的各分量表及全量表的相关系数在0.54～0.71之间，与高考总分的相关系数为0.45，这些为其效度提供了支持。

三、认知能力测验

认知能力测验是由桑代克等美国心理学家于1968年—1972年间编制成功。该测验由四个不同部分组成：

初级型：适用于小学低年级儿童。使用图片材料和口头指导语，包含4个分测验：口头、词汇、关系概念、多重智力和数量概念。

文字测验：适用于小学四年级以上。由词汇、句子填充、词语分类、词语类推4个分测验组成。

数量测验：适用于小学四年级以上。由数的大小比较，数列补充和建立关系等式3个分测验组成。

非文字测验：适用于小学四年级以上。由图形分类，图形推理和图形综合3个分测验组成。

所有测验的题目均由易至难排列，每个测验均有几套不同水平的题目，以便对智力成熟水平不同的人提供适当难度的测验，结果以离差智商、百分等级、标准九分数等解释。

认知能力测验具有相当详细的信度和效度资料，表明其各部分测验的再测信度系数在0.72～0.95之间；同时，它对学业成就、工作成就、职业类型等有相当的预测能力。

认知能力测验是一个应用相当广泛的团体智力测验，在实践中显示了较高的应用价值，只是至今未有中文修订本出现，因而在国内该测验只供研究使用。

练习与思考

1. 试谈智力测验的效度问题。
2. 智力测验的功能是什么?
3. 为什么说“智力测验面前并非人人平等”?
4. * 试析智力测验存在的合理性。
5. * 试析智力的形式与发展过程中的遗传与环境的关系。

第十三章　能力测验（下）

本章提要：

- 能力倾向测验的性能与编制、典型能力倾向测验的性能
- 特殊能力测验的作用、典型特殊能力测验介绍
- 创造力测验的理论、典型创造力测验介绍

第一节　能力倾向测验

一、能力倾向测验的产生

（一）理论的支持

在智力的心理测量学的理论研究中，由斯皮尔曼提出的一般智力因素得到了充分的重视以及普遍的认可，因而在此基础上产生了传统的智力测验并且广泛地应用于社会之中。但是，关于智力实质的理论研究和争议并未因此而停止。以瑟斯顿和吉尔福特为首的智力多因素论者虽然最后不得不承认智力G因素的存在，但他们始终强调的是构成智力的多种基本能力因素，并视其为智力结构中的核心因素群，认为应从这些不同的能力因素人手评估智力。在弗农的智力层次理论中，智力被作了进一步的细分，呈树状结构，智力的一般因素被分为几个主要的能力因素，最后再细分为众多的基本能力因素。在这样的智力结构理论的发展过程中，因素分析方法的发展和应用起了决定性作用，它使得智力所包含的各种不同能力因素能够被辨别，分类和定义，进而使得对这些能力因素有针对性地测量变为可能。

（二）实践的推动

传统的智力测验一经产生，便被迅速且广泛地应用于社会各个方面，在对个体差异的测量、尤其是对个体的分类和预测

上显示了一定的价值。但正如我们在上一章中所论述的，人们在使用智力测验的过程中，也逐渐发现其缺陷与不足之处，其中之一便是对智力测验结果的单一分数的解释问题。从智力测验所依据的理论基础来看，这一分数表明的是个体在一般智力上的差异，但在实际实施与解释中，人们发现即使某些被试得到同样的智商分数，他们在不同的题目类型中所表现出来的成绩却是不一致的。因而，要想在更精确的意义上来解释个体的能力差异，传统智力测验显然是不够的。

随着社会的发展，学校专业与工作职业的选择和咨询，以及人事工作的科学管理逐渐普及。作为心理学家和教育学家，他们所关心的实际问题之一，就是引导青年人选择和从事他们自己所喜爱的并将会有所作为的专业与工作；作为人事管理人员，他们首要的任务就是选拔出对即将担任的工作有兴趣并且完全能够胜任的人，并且将每个人安置在他最有可能发挥特长的岗位上；而对于每个人来说，他们也希望在面临求学或就业选择时，能够清楚地了解自己在不同的能力因素方面的优劣程度，从而能够最有成效地决定自己的发展方向。在这种种决策过程中，能力测验将是最主要的辅助工具之一。因此，社会需要能力测验的呼声很高，传统的智力测验在此却显得势单力薄：许多研究表明，不同性质的工作要求不同的知识和能力专长。这似乎是显而易见的事实，会计需要快捷的计算能力，律师需要领会和运用语言的能力，美术家需要空间关系的知觉和想象能力，机器修配工需要机械操作能力，等等。那么，要想在人员和工作之间作出最佳匹配，使得物尽其材，人尽其用，就必须清楚每个工作所要求的主要能力因素以及每个人员所具备的主要能力素质。此时，运用传统智力测验显然不能完美地解决问题，它只能就一个单一而笼统的智商分来解释和比较个体间差异，却不能对此差异内部的一些现象做更进一步的客观

判断，更无法比较不同能力因素上的个体内差异。因此，它既不能为个体提供自身在不同能力上的水平分布状况，也不能帮助人事管理部门有效选择和安置在工作所需的特定能力上水平相对最高的人员。

由于社会迫切需要能够辨别和判断在不同能力因素上的个体间以及个体内差异，而当时现有的能力测验——智力测验又无法满足这一要求，同时，因素分析方法又使一般智力中所蕴含的各种不同能力因素的辨别分类成为可能，能力倾向测验应运而生。

1941年，瑟斯顿在自己的智力理论支持下，编制并且发表了第一个能力倾向成套测验——基本心理能力测验(PMA)，主要测量五种能力因素：言语能力、数的能力、知觉速度、推理能力和空间关系认知能力，目的是为了了解和预测学生在各门课程中的学习情况。由于该测验存在许多技术缺陷，因而现在较少应用，但它具有重要的历史意义，为此后发展起来的各种能力倾向测验的先驱。

在近几十年里，能力倾向测验得到迅速发展，新测验层出不穷，并且大量地应用于社会，成为人员选拔与安置等决策的有效辅助工具。

二、能力倾向测验的特点

结合能力倾向测验产生的理论及实践背景考虑，其特点主要表现于以下几个方面：

（一）测验的目的

成就测验的目的与能力倾向测验不同。所谓成就，是指个体经过一定的教学或训练后所掌握的知识水平或所达到的能力水平，它针对于特定的学习经验，强调个体目前已经达到的水平。换言之，成就测验是对个体过去学习经验的总结。

能力倾向指的是个体在不同能力因素上潜在的优劣倾向。当能力倾向测验首次出现时，能力倾向被定义为不是建立在经验之上的、特殊的、天生或遗传的能力。现在看来，这一定义似乎有失偏颇。严格地说，任何能力都在一定程度上既依赖于遗传的潜能，又依赖于生活环境中所积累的经验，能力倾向同样如此。然而，能力倾向虽也依赖于个体的生活经验，但并不直接依赖于专门的教学或训练。卡洛尔（J. B. Carroll）等人的一些研究表明，具体的课程教学或知识技能训练可以显著改善成就测验的成绩，面对能力倾向测验的成绩却没有影响。因而，能力倾向测验的目的不在于总结过去，而在于预测将来，即预测个体在将来的学习或工作中可能达到的成功程度。如某人的测验结果表明他在逻辑推理能力上有明显的优势，我们可以预测此人将来在理科课程的学习中可能取得较好的成绩。

虽然智力测验的主要目的也在于预测，但由于其所依据的理论基础与能力倾向测验不同，它所作出的预测比较笼统，针对性较弱。而能力倾向测验预测的目的性更强，它试图说明个体在多种能力上的潜在优势，并进而和专业或工作所需结合起来。

（二）测验的编制

能力倾向测验一般同时测量几种能力因素，以分测验形式组成，每个分测验针对一种能力，每个测验应该是独立的，并

且，各分测验间的相关要尽可能低。测验的内容涉及广泛，不像成就测验那样具有明确限定的内容范围，且较少涉及与学校习得知识有关的内容。

由于各分测验的结果不仅要在个体之间进行比较，而且还要在个体内部进行比较，因此各分测验必须使用相同的常模样本，且应具有较高的信度。

（三）测验结果的解释

被试在能力倾向测验上可以得到若干测验分数。这些分数既可表明不同被试在每一分测验所测能力上的相对位置水平，又可表明同一被试在所测各能力上的相对优劣状况。一般能力倾向测验往往会用能力剖面图来呈现个体内差异。

由于不同工作一般需要不同的能力特长，但又往往不只需要一种能力，因此，当用能力倾向测验预测个体在某工作上的可能成功程度时，需要解决各分测验的分数组合问题，即如何确定各种能力因素（各分测验分数）的理想权重。针对不同工作，不同能力因素的权重应有变化。一般采用多重回归模式解决这一问题。

三、具体能力倾向测验介绍

（一）学术能力倾向测验（SAT）

学术能力倾向测验（Scholastic Aptitude Test，以下简称SAT）相当于我国的高考，是大学录寻取新生的一项主要参考依据，每年在美国和世界各地举行多次．SAT由美国教育测

验服务中心主持试题编制和试卷分析等工作，几十年来技术不断改进，如今已属于技术上最完备的测验之一，每一新试卷都已达到了高度的标准化。

SAT测量的目的不在于总结学生在中学时学到多少知识，而在于预测学生是否具备大学学习和研究的能力，以及倾向于在哪些专业领域更具优势，因此SAT筛选题目的主要依据是预测外在效标的有效性，测验材料一般避免过多依赖具体的知识和教学经验，而是迁移到各种广泛情境的技能和材料上，学生则必须把他的知识和能力应用到新异的情境。

SAT包括两部分内容：语言和数学。语言部分包括反义词、句子填充、类比推理、阅读理解等内容，考查学生在词汇量、阅读理解、逻辑思维、以及作出判断和结论的能力。数学部分包括算术、代数和几何等内容，考查学生在数学运算、推理能力以及应用数学概念与知识解决实际问题的能力。

SAT题型皆为多重选择题，有四五个选择项，测验时限为3个小时，测验结果包括语言和数学两个分数，没有合成分。一般大学不会公布录取分数线，因为美国录取新生不仅参照SAT成绩，同时还要结合学生的中学成绩单、教师推荐信、所在中学的相对水平以及学生的性格、兴趣和特长等多方面资料来综合考虑。

（二）分辨能力倾向测验（DAT）

分辨能力倾向测验（Differential Aptitude Test，以下简称DAT）是由美国心理公司于1947年初版，并于1963年和1972年两次修订和进一步完善，是应用最广泛的成套能力倾向测验之一，主要适用于初中和高中学生的教育咨询及就业指导。

DAT包括8个分测验、单独施测并单独记分：

1. 言语推理

测验项目类型为类比推理，每题提供5对备选答案，内容涉及历史、地理、文学、科学等多方面知识，目的在于测量和评价个体的言语理解与抽象概括以及作建设性思考的能力，从而进一步预测个体是否适宜从事以复杂的言语关系及概念为主的学科或职业，如高深的科学研究工作等。

2. 数的能力

测验项目类型为计算题，不过题目具有一定的复杂性，并不是只反映计算的熟练程度，还需要对数目关系的理解能力以及处理数目概念的灵活性。测量目的在于评估个体对数目进行推理，思考数量关系以及明智地处理数量材料的能力，进而对个体在教育或职业方面的选择与发展作出预测，如：教育方面，可用于预测数理化、工程等学科；职业方面，可用于预测统计工作者、工艺制作者以及与自然科学有关的各种职业。

3. 抽象推理

测验项目是非文字材料，呈现的是一组组成一定联系或按次序排列的问题图形，要求被试找出可使这种排列连续下去的图形，作答关键在于找出每组图形变化的原则或规律，和言语推理并不相同。不过对于言语方面不能沟通的被试，本测验分数可以校正在言语推理测验的得分。

4. 文书速度与准确性

测验要求被试首先在测验本上选出画了记号的一个符号组合，然后在答案纸上找出相同的一个组合。测验项目所提供的情境和一些实际的文书工作比较相近，目标在于测量对简单知觉工作的知觉速度、短时记忆和反应速度，是DAT中唯一以速度为主的测验，对于档案或资料整理及管理等方面工作具有一定的预测意义。

5. 机械推理

测验项目设计一些机械装置或情景，要求被试指出哪种选择符合情景，测量对表现于熟悉情境中的机械和物理原理的理解力，但被试是否受过物理学的正式训练对测验结果影响不大。凡含有普通物理原则的课程及职业，如物理或机械技术等课程以及木工、机工、机器装配与维修等工作，都需要一定的机械推理能力。

本测验的结果存在显著的性别差异，女生的分数普遍低于男生。

6. 空间关系

测验项目要求被试能在心理上操纵三维空间，即能够对所显示的平面图在想象中从不同方位进行转换和折叠，测量个体经由视觉想象处理具体材料的能力。很多专业或职业需要这种空间知觉能力，如美术、建筑、服装设计等。

7. 语言运用：拼写

测验列出了一个单词表，其中有些单词有拼写错误，被试必须指出每个单词的拼写正误。

8. 语言运用：文法

测验项目由若干句子组成，每个句子被记号划分为几个部分，要求被试从语法或修辞等角度找到错误或不合理的那一部分。

测验 7 和 8 的内容和形式更近乎于成就测验，但由于它们测量了诸如速记、秘书、新闻、广告等若干专业或职业活动中所必须的基本技能，因而被纳入 DAT 中。由于这两个测验之间相关较低，因而分别计分。

除了文书速度与准确性测验之外，DAT 的分测验基本上是能力测验，可团体施测，其时间限制在多数情况下为 30 分

钟，成套测验总测时间大约为 $5 \sim 5\frac{1}{2}$ 小时，一般至少分两次进行。每一分测验都有年级百分常模（从八年级到十二年级），语言推理和数字能力的组合也有常模，该常模可用于评价一般的学术能力倾向。根据常模将每一个体在测验上的原始分数转化为百分位数后，可绘制个人能力倾向剖面图。

DAT 的能力剖面图既可直观提供个人在 8 种能力倾向上的内部差异，又可表明个人在每种能力倾向上相对于同年级团体的相对位置。因而，该剖面图可帮助学生了解自己的长处和弱点，从而更能了解自己在某些学科学得好或不好的原因，进而有效选择自己今后的教育和职业方向。并且，学生也可从自己的能力剖面图中发现自己以前未曾认识到的或被低估的潜力，从而激发较强的成就动机。对于学校而言，当他们录取学生时，DAT 可为他们提供每个申请者在多方面能力因素上的表现。此时，学校可根据自身的专业及课程设置状况，建立起一组临界分数，作为录取学生的标准，也可进一步用于安排已录取学生的专业。

可见，DAT 被广泛用于教育领域中对于学生将来学术成就的预测方面，这就要求 DAT 中不同的测验对不同的学科的预测是有效的。DAT 手册中提供了丰富的效标关联效度资料，表明了每个分测验对不同的学科的预测力是有差异的，如言语推理测验的结果可以较好地预测英语、社会科学等学科成绩，但对于数学的预测效果较差；又如机械推理在对自然科学学科，打字和工艺方面的成绩预测上比对数学、英语、社会科学学科等的预测更为有效。这些预测上的差异说明利用 DAT 的不同分测验来预测学生将来的学术成就是有效的。同时，效度资料显示，言语推理和数学能力的合成分数对于所有学科都有较好的预测作用，因而这一分数可被看作一般学习能力倾向的

体现。

最后需要说明的是，DAT 虽然为个人或学校或咨询者提供了比较丰富的能力水平资料，从而为决策提供了依据，但仅仅以此为依据尚不足以做出准确决策，而是应同时结合其他资料，如兴趣测验结果、学业成绩、个人志愿、家庭背景等进行综合考虑。

（三）一般能力倾向成套测验（GATB）

一般能力倾向成套测验（General Aptitude Tests Battery，以下简称 GATB）是本世纪 40 年代由美国劳工部就业保险局设计而成的综合式职业性向测验。GATB 是在对早先为某些工作而准备的 50 多种测验进行因素分析的基础上编制而成的，包括 12 个分测验，可用于测量 9 种能力倾向因素：

1. 一般智力（G）

指掌握基本原理、原则以及做出推理、判断的能力，它常与学业成绩有密切相关，可由词汇、算术推理和空间关系 3 个分测验相结合测量而得。

2. 言语能力（V）

指能了解文字的意义，掌握字与字之间关系并能有效使用文字的能力，可由词汇测验来测量。该测验要求被试从 4 个一组的词汇中找出成对的同义词或反义词。

3. 数的能力（N）

指能正确而迅速地作加减乘除运算，并能利用算术知识解决实际问题的能力，由计算和算术推理两个测验相结合测量而得。这两个测验分别由四则运算题和应用题组成。

4. 空间关系理解力（S）

指能在心理上将平面图形转换为具有三维空间关系的立体图形，并能从不同角度认识同一物体的能力，由空间关系测验

进行测量，测验项目呈现一个平面图形和四个备选的三维图形，要求被试判断哪一个三维图形是由给出的平面图形折叠而来。

5. 形状知觉能力（P）

指能觉察到实物或图形的细节、能对图形的外形与明暗上的差异或线条在长宽上的细小差异作正确的比较和辨别的能力，由工具辨认测验和图形配对测验联合测量。这两个测验要求从备选项中选择出与给定工具或图形相同的答案。

6. 文书知觉能力（Q）

指能觉察文字、符号、表格上细微差异以及能快速校对文字、数目、符号以避免抄写或计算错误的能力，可由文字校对测验来测量。此测验项目类型类似于工具辨认和图形配对测验，只是测验材料由文字取代了实物和图形。

7. 动作协调能力

指能使手指之间和手眼之间相互协调配合，作出快速且精确的细微动作的能力，可由画记测验测量。该测验要求被试在一系列方格中，用铅笔作出特定的记号。

8. 手指灵巧性

指能灵活运用手指、以双手手指快速且精确地分解或组合小物体的能力，可由装配测验和拆卸测验相结合进行测量。这两个测验使用同一装置：一块板的两头各有 50 个孔，在其中一头的每一个孔中放有一个小铆钉；另外，在一个转轴里放了一迭垫圈。在装配测试中，要求被试用一只手拿起一枚铆钉，另一只手拿起一个垫圈，并把垫圈放在铆钉上，然后把它们装配到板的另一头相应的孔上，时间限制在 90 秒钟内。在拆卸测试中，被试要拆掉装配好的铆钉和垫圈，把它们放回原处。结果以装配或拆卸的件数计分。

9. 手部灵巧性

指能灵活运用手腕、手肘，将物体作快速且精确的移动或转动的能力。该能力由两个拼板测验来评定：在第一个测验中，被试用双手把置于一块拼板各个孔内的栓子移到另一块拼板上去；在第二个测验中，被试用他认为比较灵活的那只手从拼板上拔起一个栓子，在手中旋转 180 度，再把这个栓子的另一端重新插到孔内。

GATB 的 12 个分测验中，既有纸笔测验，也有操作测验。因此施测时纸笔部分可团体进行，而操作部分则必须个别施行。与 DAT 相比，GATB 纳入了 DAT 所没有的形状知觉测验及几种运动能力测验，因而，GATB 比较注重于实际操作，而且多个分测验更倾向于速度测验而非能力测验。但是尽管如此，在 DAT 和 GATB 中对应的因素之间还是密切相关，如言语、数字、空间关系、文书等因素在 DAT 和 GATB 之间的相关系数非常显著，其值在 0.57～0.74 之间。

个体完成 GATB 的 12 个分测验后，可得 9 个原始分数，分别针对于上述 9 种能力因素。测验选用一般在职人员为常模团体建立常模，个体在测验中的原始分数根据此常模转换成平均数为 100，标准差为 20 的标准分数，然后可绘制成能力剖面图。从图中可以直观地看到个体内部在 9 种能力因素上所表现出来的优劣倾向，又可比较和判断个体相对于一般在职人员在 9 种能力因素上的相对水平，因而对于个体的就业指导、决策以及人事部门的人员甄选和录用具有相当高的辅助价值。

由于不同的能力因素在不同的职业中所显示出来的重要性不同，因此在职业辅导或人员甄选时，除了解个人在各方面能力上的优劣之外，还必须了解各种职业最需要什么样的能力，以及在所需要的能力上水平达到什么程度才能胜任相应工作，这样才能更准确地判断某个人是否适合某种职业。在这种考虑

之下，GATB选用了若干种职业，从相应职业的在职人员中选取代表性样本作为常模团体，建立了若干个个别职业常模。将个人的9种能力因素标准分与某一个别职业常模所要求的能力因素的切割分数相比，可能的评价为高、中、低三种：被评为“高”，表示此人的能力符合且超过该职业的合格员工，在该职业中成功的可能性很大；被评为“中”，表示此人的能力接近该职业的合格员工，可以胜任该职业；被评为“低”，则表示此人从事该职业的成功可能性较小，应考虑从事其他更能发挥能力的职业。

为了更好地应用于管理和选拔工作，美国职业介绍服务机构以工作分析为基础，把各种职业进行分类，总共设置了36个职业群，并建立了相应的常模，每个职业群常模均规定了自己的分数线模式。据此常模，可以判断个体是否适合从事某一类职业，以及获得成功的可能性有多大。分数比较和解释的方法与个别职业常模类似。

GATB手册提供了为数可观的不同职业工人在GATB测验结果上的差异的数据，以及不同分测验与不同工作之间的相关数据。这些数据资料为GATB的效标关联效度提供了证据。

将GATB的结果应用于实际的职业指导和人事工作时，同样应避免完全地和单纯地依赖此结果，因为它只是必要条件而非充分条件。

第二节　特殊能力测验

能力倾向测验可以从不同能力因素上来评估个体，因而被

广泛应用于专业、职业指导与人事管理工作之中。但是，它们在使用过程中也逐渐显现出一些弱点：当个人已有强烈的志愿去从事某类专业或职业时，他希望能有一种测验可以针对此类专业或职业所需要的特殊能力进行测量，从而可以评估自己将来在此类专业或职业领域成功的可能性大小，并据此调整或确定志愿；当某人事部门需要招收特定工作岗位的工作人员，或某专业学校需要招收学生，或某人眼前正有某一工作或某一学校的选择机会；当特定人员与特定工作或学校之间意向性很强时，无论从个人角度，还是从校方或雇主角度来考虑，都希望能有一种适合自己的专业性较强的特殊能力测验，并且，能根据测验结果来评估与判断是否录取某人或是否选择某工作(或学校)。成套的能力倾向测验在此便显得有些累赘。虽然也有人主张在这种情况下可以不对被试施行整套测验，而是有针对性地选择施测某几个分测验，从而简化测量程度并提高测量效率。

但是，这样的做法并不值得提倡。因为：一方面，包含在成套测验中的每一分测验尽管目的在于测量各种能力因素，但由于时间的限制，每个分测验不可能编得很长，所包容的内容也不可能很多，这样一来，其题目取样便受到了相当程度的限制。因此，将每个分测验作为单独的测验来使用便会显得不够充分。另一方面，成套能力倾向测验虽然往往包含了多种能力因素的测量，但一个测验不可能涵盖所有能力因素，其所制定的职业常模也不可能涵盖所有职业。因此，对于某些专业领域的人员选拔，能力倾向测验中若无针对相应能力的分测验，此时便无法提供任何帮助。

出于以上考虑和需要，特殊能力测验应运而生。现有的常用的特殊能力测验一般针对于一种特殊能力所包含的各方面因素进行测量，测验性质介于成就测验和能力倾向测验之间，其

内容与相应的专业或职业训练的重点是一致的，而测量目的既想了解个体在此专业领域的既有水平，又想预测个体今后在此专业领域成功的可能性。常见的特殊能力测验主要有音乐能力测验、美术能力测验和机械能力测验。

一、音乐能力测验

当我们涉及音乐和美术这样的领域时，一般的学习能力倾向测验便显得有些不合适了，它们对这些学科的成绩是难以预测的。就音乐能力而言，它包括各种感觉辨别力，如音调辨别力、音量辨别力、音程长短的辨别力等，也包括对音乐题材中较复杂的音乐关系的理解，如音程关系、曲调类型和音的构成等，同时还包括关于曲调、节奏、格调、强弱等的艺术判断力。能力倾向测验基本上无法涉及这类的能力范围。爱荷华大学的西肖尔（Carl Seashor）及其同事对音乐能力进行了开创性研究，结果产生了最早的音乐能力测验。

（一）西肖尔音乐才能测验

西肖尔音乐才能测验（Seashor Measures of Musical Talents）是一组最充分地调查音乐才能的测验，它以一系列音乐调式或音乐符作为刺激材料，主要测定以下一些简单的感官辨别力：

（1）音调辨别力：判断两个调子哪一个较高。

（2）音量辨别力：判断两个声音哪一个较响。

（3）时间音程辨别力：判断两个音程哪一个较长。

（4）节奏判断力：判断两个节奏是否相同。

(5) 音色判断力：判断两个音色哪一种较悦耳。

(6) 音调记忆力：判断两首曲调是否相同。

本测验适用于小学生到成人，共需1小时左右完成。测验手册中有明确的信度分析，但效度证据还不够。

效度问题使西肖尔音乐才能测验受到批评与怀疑，并且，它所选择的刺激材料被认为远离了真正的音乐题材，因而引致的争议更大。后期的音乐能力测验便趋向于采用更复杂的内容。

(二) 音乐能力测验图

音乐能力测验图（Musical Aptitude Profile）是由弋登（E.Gordon）1965年编制而成的。它以真正的音乐题材为材料，包括250个原版的小提琴和大提琴选段。

测验首先包括若干个对音乐理解力的测量项目，要求被试分别以旋律、和声、速度和拍子为基础来比较和判断两小段音乐之间相异或是相同。其后，测验提供的是3个分测验：

(1) T测验——音调形象（旋律、和声）。在该测验的音乐表达方式上有两种演奏方法，让被试判断异同。

(2) R测验——节奏形象（速度、节拍）。演奏有两个结尾，亦要求被试判断异同。

(3) S测验——音乐感受（短句、平衡、风格等）。要求被试判断两段音乐哪个更有韵味。

本测验具有相当吸引人的信度资料：每个分测验的信度均在0.80左右，合成分的信度在0.90以上。测验的效度也有一定的证据：测验结果与教师评定之间的相关在0.64～0.97之间；弋登对200多名学生进行为期三年的追踪研究，发现测验成绩可对学生在专业音乐训练上的成绩作出较好的预测，并且所进行的音乐训练的时间越长，这个测验作出的预测就越准

确。因此，该测验在技术上比西肖尔测验更为完善。

二、美术能力测验

关于美术能力的判断标准是很难确定的，因而寻找可靠的标准并据此编制可靠的用以测量美术能力的测验也是很难的。不过，尽管如此，仍有许多有关美术能力的测验产生。

（一）梅尔美术判断力测验

梅尔美术判断力测验（Meier Art Tests）的每一个项目都由两幅美术图片组成，一幅是公认的杰作，另一幅是在某些方面（平衡、比例、明暗等）对此杰作稍有歪曲的作品。测验指导语中简要指出了两幅图片的差异，要求被试在这两幅片中选出更好的一幅。

测验分半信度从 0.70～0.84，但评分者信度不高。在测验上得分的高低，表明被试对于美术作品的鉴赏能力，这可以说是美术能力中最基本的部分，是从事各种与美术有关的学习或工作必备的，一个缺乏审美能力的人，最多只能成为一个普通的艺术工匠，永远不可能成为一个真正的美术家。当然，具有很高审美能力的人，也不一定就必然成为一个好的美术家。因而，本测验只是考察美术能力和预测美术成就的一个必要条件，但非充分条件。

（二）格雷福斯图案判断测验

格雷福斯图案判断测验（Graves Design Judgment Test）的取材不再是名家杰作，而只是一些二维或三维的抽象图案，

每一项目包括 2～3 个同一图案的变式，它们在整体性、平衡性和对称性等方面有所区别，要求被试判断出哪一图案最好。

测验分半信度为 0.80～0.90。测验结果可表明被试对美术一般的基本原理的认识和反应，从而证实他对美学知觉和判断的标准。该测验没有提供足够的效度证据。

（三）霍恩美术能力问卷

霍恩美术能力问卷（Horn Art Aptitude Inventory）由两部分组成：第一部分要求被试画出 20 种常见的物体和几何图形；第二部分要求被试在长方框内给定的圆点和线条基础上作画。

该测验是操作型测验，可以让被试显示一般美术记忆和技巧以及美术想象和创造力。不过测验评分的主观性太强，多少限制了测验的应用。

美术能力测验一般可以成功地将美术学生或美术工作者和其他人员区别开来。对美术学生的研究表明美术能力测验的成绩对于他们后来在美术学校所取得的成就是一个很好的预测。至于这些测验对未受过美术训练的学生能起多大作用，仍待进一步研究。

三、机械能力测验

大多数工业职业需要一定的机械能力，因此对于个人在工业职业方面的可能成就的预测就需要机械能力测验的参与。机械能力包含了许多成分，如运动能力、空间知觉能力、机械推理等等。现有机械能力测验往往只就某一成分进行测量，并且主要集中于对空间关系能力和机械知识、理解及推理能力这样

两个方面加以测量。

(一) 空间关系测验

明尼苏达大学的帕特森（D.G.Paterson）及其同事对机械能力作了严格的分析，并编制出3个有关的测验：

1. 明尼苏达机械拼合测验（Minnesota Mechanical Assembly Test）

这是一个操作测验，要求被试拼排随机排放的机械物体，主要测量动作敏捷性和空间知觉。

2. 明尼苏达空间关系测验（Minnesota Spatial Relations Test）

测验材料是四块带有各种几何形状凹陷的板，两套随机放置的具有各种几何形状的木块，要求被试尽快将木块放入相应几何形状的板中。测验主要考察被试对空间关系的知觉速度，具有较高的信度和一定的效度。

3. 明尼苏达书面形式拼板测验（Minnesota Paper Form-board Test）

测验以纸笔形式进行：采用多重选择题，每题均由被分解开来的几个几何图形组成，要求被试从备选答案中选出由这几个几何图形拼合起来的整体图形。该测验具有较高的信度。并且在预测机械操作及包装检验等工业职业的实际成就上显示一定的价值。

(二) 机械理解能力测验

机械理解能力是指理解实际生活情境中的机械原理的能力，一般需要一定的机械知识。本耐特机械理解测验（Bennett Mechanical Compredension Test）是对此能力进行测量的最常见的工具之一。

本耐特等人将大量的日常生活情境引入测验材料，每题皆以图示，图旁附有一个简短的问题。比如图示中两个人用一长条木板抬一重物，重物距一人近而距另一人远，问“哪个人负重更大”。

由于在机械能力上存在明显的性别差异，因此测验根据性别分别建立常模，被试在测验上的得分与同性别的常模进行比较和解释。测验具有相当高的信度，并且，该测验结果与若干工业职业成就的相关研究为测验提供了较好的效度证据，因而，此测验在军队和企业里以及一些专业学校（如工程学校）里都得到了广泛的应用。

第三节　创造力测验

文明的历史，基本上乃是人类创造能力的记载。创造力是推动人类社会发展的原动力，作为一种特殊的而又不容忽视的能力因素，创造力受到了心理学家们的注意，关于创造力的探讨和研究在近几十年来也成为心理学的热门话题之一。

一、发散思维研究与创造力测验

创造力研究被纳入科学研究轨道之后，在很长一段时间，关于创造力的探讨停留在思辩阶段，高尔顿将之归结于遗传，以弗洛伊德为首的心理分析学派将之归结于无意识过程，格式

塔学派又将之归结于顿悟，等等。由于研究方法和工具的不足，研究者们各执己见，意见纷纭。1950 年，吉尔福特在美国心理学年会上作了题为《创造性》的著名演讲，此后，许多创造力研究者皆遵循他的思路继续研究。

吉尔福特在智力结构的研究中引入因素分析方法，由此提出了他的智力三维结构模型（见上章）。在此模型中，他发现智力操作中存在聚合与发散两种不同类型的思维：聚合思维是指利用已有的知识经验或传统方法来解决问题的一种有方向、有范围、有条理、有组织的思维方式；而发散思维则是既无一定方向又无一定范围的由已知探索未知的思维方式。

吉尔福特还认为发散思维在行为上主要表现出 3 种特性：

①流畅性：面对智力任务能在短时间内作出迅速而众多的反应。

②变通性：思维灵活多变，触类旁通，不受传统思维或心理定势的影响，能多方位地思考与解决问题。

③独特性：对事物能表现出不同寻常的新颖见解。这 3 种特性相互联系，变通性建立在流畅性基础之上，独特性又建立在变通性与流畅性的基础之上，因为只有反应数量众多，才有可能反应角度多样化，进而才有可能出现新视角，新观点。

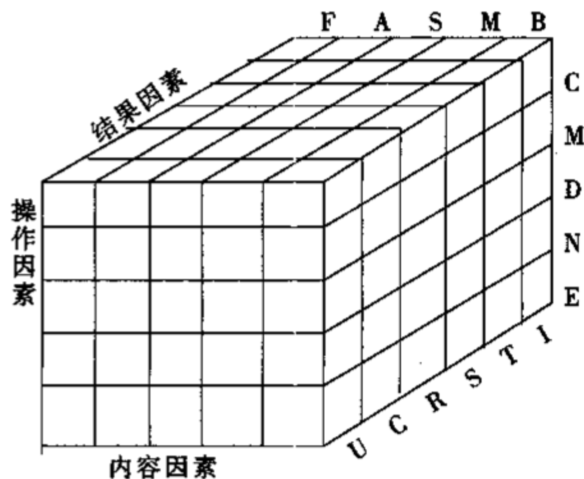
吉尔福特将发散思维的特性视为人的创造性活动的特性，并因此将创造力定义为发散思维的能力，即对规定的刺激产生大量的、变化多端而又独特的反应的能力。他进而指出：现有的传统的智力测验一般注重于聚合思维的测量，测验项目通常要求被试从给定的若干备选答案中选出一个，评分则以固定的正确答案为标准，并不鼓励被试作出多样化的与众不同的反应，因此，被试的创造力在智力测验中无法得到充分的反映。然而，随着创造力研究的深入以及社会发展对于创造性人才的需求日益增加，关于创造力的测量已经逐渐具备了理论上的可

能性和实际上的必要性，因而势在必行。

吉尔福特关于智力测验注重聚合思维而忽视发散思维的评论得到很多学者的共鸣，并且，他视发散思维为创造力之核心的观点也为很多研究者所接受。因此，目前常见的、有一定影响力的创造力测验基本上是沿循吉尔福特的理论观点编制而成的。

二、吉尔福特发散思维测验

吉尔福特在长期的研究中设计出大量的测验对发散思维进行测量。这些测验将他关于创造力的定义和他关于智力结构的阐述结合起来：视创造力为发散思维能力，发散思维又是智力三维结构中操作维度所包含的五个因素之一；而作为操作因素，发散思维又可以与智力结构中的5种内容因素，以及6种结果因素之间组合出30种心理能力因素，如图13.1所示：



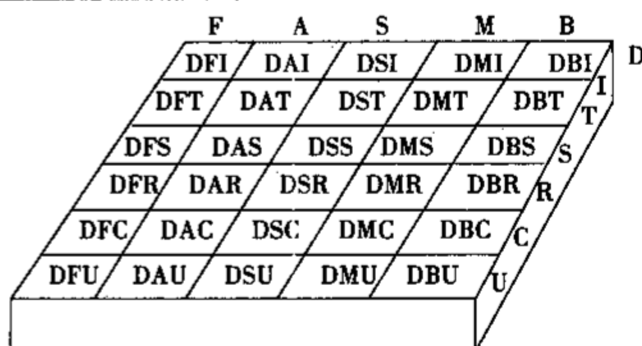


图 13.1 吉尔福特发散思维测验的理论模型

图中上半部分即智力三维结构模型，下半部分则将发散思维部分从模型中抽取出来，并且已标明与发散思维相关联的30种心理能力因素及其位置（图中各字母符号所代表的含义请参阅上章）。

吉尔福特力图选择合适的方法来测量图 13.1 中所示的 30 种能力因素，但最后只编制出 14 个分测验，针对其中 11 种能力因素进行测量。

(1) 词语流畅：写出包含某一指定字母的词，测量 DSU 因素。

(2) 观念流畅：列举属于某一种类的事物的名称，测量 DMU 因素。

(3) 联想流畅：列举近义词，测量 DMR 因素。

(4) 表达流畅：给定 4 个字母，要求写出所有可能的由 4 个以给定字母开头的词组成的句子，测量 DMS 因素。

(5) 多项用途：列举指定物体的各种不同寻常的用处，测量 DMC 因素。

(6) 解释比喻：用不同方式完成一个比喻句，测量 DMS 因素。

（7）效用测验：列举某物的所有可能用途，测量 DMU、DMC 两因素。

（8）故事命题：写出一个短故事情节的所有合适标题，测量 DMU、DMT 两种因素。

（9）推想结果：列举一个假设事件的所有可能结果，测量 DMU、DMT 两因素。

（10）职业象征：列举一个给定的符号或物体所象征的可能职业，测量 DMI 因素。

（11）图形组合：仅仅使用一组给定的几何图形，画出指定的物品，测量 DFS 因素。

（12）绘图：以给定的简单图形为基础，绘出尽可能多的可辨认物体的草图，测量 DFU 因素。

（13）火柴问题：移动指定数量的火柴，形成特定数目的方形或三角形，测量 DFT 因素。

（14）装饰：以尽可能多的方法来修饰一般物体的轮廓图，测量 DFI 因素。

测验一般适用于初中水平以上的人，从思维的流畅性、变通性和独特性 3 方面进行评分。分半信度在 0.60～0.90 之间，测验手册中报告了每个测验的因素效度，但缺乏效标关联效度的数据资料。

三、托伦斯创造性思维测验

托伦斯创造性思维测验（Torrance Test of Creative thinking）是在吉尔福特的智力理论及其发散思维测验基础上编制而成的，目的是从流畅性、变通性、独特性和精确性 4 个方面

评估个体的创造性思维能力。测验共分两套，每套都有两个复本。

（一）言语的创造性思维测验

这一套测验包括7项活动：

（1）发问：呈现一张图画，要求列举为了解图中之事而需要询问的所有问题。

（2）猜测原因：列举图中之事发生的所有可能原因。

（3）猜测结果：列举图中之事的所有可能后果。

（4）产品改进：对给定玩具提出改进意见。

（5）非凡用途：列举某物不同寻常的可能用途。

（6）不平凡的疑问：对活动5中所示物体提出不同寻常的疑问。

（7）推想结果：列举一种假想事件的所有可能后果。

（二）图形的创造性思维测验

此套测验包括3项活动：

（1）建构图画：以明亮的彩色曲线为起点，建构一幅故事画。

（2）完成图画：利用所给的少量不规则的线条画出物体的略图。

（3）平行线条绘图：利用成对的平行线条绘出尽可能多的不同的图形（复本中以圆代替平行线）。

测验结果得到流畅性、变通性、独特性和精确性4个分数。在判断一个人的创造性思维能力时，必须4个分数综合起来分析，而不能根据某一孤立的分数进行推断。测验的分半信度和复本信度在0.60～0.93之间，但没有充分的效度证据。

创造力测验的产生使得创造力研究更加深入，但也带来了更大的争议。一方面是理论上的争议：究竟发散思维是否就是

创造力的核心？有些研究者认为聚合思维对创造性活动同样重要，一个真正具有创造力的人，不仅要有变通而独特的思维，而且也需要有效地选择、评价与综合的思维能力，这样才能将他的奇思怪想与现实情境结合起来作出成就，否则空想永远是空想。也有的人认为应该将个性因素考虑到创造力中去，一个高创造力的人应该具有好奇心、独立自主性、自信心和冒险精神等个性特征，创造力测验应从这一角度入手。总之，创造力包罗万象，应从多维度进行系统而综合的研究与评估，单从发散思维角度来测量显然是不够的。

另一方面是创造力测验本身的一些弱点限制了它们的应用：①创造力测验的评分较为复杂，虽然测验手册上有详细的评分准则，但是主观性依然很强，评分者之间的一致性程度较低，尤其是在对被试答案的独特性评分上更是见仁见智，难以统一。②测验的效度也还存在怀疑。目前常见的大多数创造力测验缺乏足够的效度证据，因而，这些测验在实际的创造性成就的预测上究竟有多大效用，目前依然值得探讨。

总之，创造力测验乃至其理论依据还处于探索阶段，其在实际预测的可靠性与有效性上都有一定的局限性，因而目前这些测验仍被视为研究的工具，而不能被施用于实际预测之中。

练习与思考

1．能力倾向测验与智力测验、成就测验之间有何不同？

2．能力倾向测验与特殊能力测验在应用上有何区别？

3．现在流行的创造力测验依据的理论基础是什么？

4*．试析社会的发展与需要在能力测验的产生与发展中的影响。

5*．试析创造力的实质与表现。

第十四章　人格测量

本章提要:

- 人格与人格测量
- 人格测量的真实性问题
- 自陈量表的编制及其特点
- 几种主要人格量表的使用
- 投射测验及其理论基础
- 罗夏克墨迹测验简介
- 主题统觉测验简介

人格测量是心理测量的一个主要组成部分，它对于在较短的时间内较为全面准确地了解一个人的人格特征，对于因材施教，对于心理异常的诊断，对于人员的选拔与任用都具有重要的参考意义。本章首先讨论人格测量的一般问题，然后分别介绍主要的人格测量工具——自陈测验和投射测验，以及它们的理论基础和使用方法。

第一节 人格测量的一般问题

一、人格与人格测量

人格（personality）是一个具有多重含义的概念，在不同的学科有着不同的意义，用在不同的场合表达不同的意思。哲学上的人格通常指人的本质属性，即人与动物所区别的那些方面。有的哲学家强调人的理性，有的哲学家强调人的自我意识，有的哲学家又强调人的理想品质。伦理学上的人格是指人的优秀品质和善良品德，类似于通常所说的道德品质。人格在法学上又是指人的权利和尊严。而在社会学上却指一个人在社会舞台上所扮演的角色。心理学家对人格的心理学含义尽管存在众多不同的看法，但在通常意义上是指一个人相对稳定的心理特征和行为倾向。在这种意义上说，人格就是中国人通常所理解的性格。正因为如此，有的研究者为了避免引起理解上的

混乱，主张将心理学上的 personality 翻译成“性格”[①]。

西方心理学家对于人格的本质持有不同的认识，对人格的研究角度不相同，因此给人格一词所下的定义也就有区别。对不同人格定义的详细讨论是普通心理学尤其是人格心理学的任务，因此不在这里详述。这里需要说明的是，各种人格定义并不是水火不容的。其中，有的是因研究的角度不同引起概念上的分歧，实际上是可以互补的；有的可能仅仅是措词上的区别，内容大致相同。

我们认为，现代西方心理学家对人格本质的理解至少在 4 个方面是一致的或基本一致的。第一，绝大多数心理学家都强调或事实上承认人格的整体性。人格虽然可能表现为各种不同的具体形式，但各种心理成分彼此交织，互相结合，组成一个整体。第二，所有心理学家都承认人格的独特性，即承认没有两个人的人格是完全相同的。第三，绝大多数心理学家都承认人格对个人行为的调节功能，即认为人的行为至少部分地决定于行为者的人格特征。第四，所有的心理学家都主张人格的相对稳定性，即认为人格对行为的调节功能具有跨时间和跨情境的特征。因此，一个人格定义无论用什么样的词语表述，只要包含了上述 4 方面的内容，就抓住了人格的实质。根据这一认识，我们把人格（或性格）理解为个人在与环境的相互作用过程中形成的相对稳定的心理特质和行为倾向的整体组织，它决定着个人行为的独特性。这个定义并不意味着完善，但它包含了人格的主要性质，也容易理解。

人格测量就是通过一定的方法，对在人的行为中起稳定的调节作用的心理特质和行为倾向进行定量分析，以便进一步预测个人未来的行为。在心理与教育测量史上，首先提倡用科学

① 沙毓英、张锋等：《中国人性格研究的理论与方法初探》，《云南师范大学学报》，1993 年第 3 期

方法测量人格的是英国学者高尔顿。早在1884年，他在《品格的测量》一文中就提出：构成我们行为的品格是一种明确的东西，所以应该加以测量。他认为通过记录心律和脉律的变化可以测量人的情绪，通过观察社会情境中人的活动可以评估人的性情、脾气等特征。他还编制了一个人格的评定量表，可以说是对人格测量技术的初步尝试。

1892年，克雷普林将联想测验用于临床诊断。其基本作法是给被试一些经过专门选择的词作为刺激词，要求受测者在听到或看到刺激词后说出他最先想到的词（反应词)。然后，通过分析受测者的反应词的内容来判断受测者的人格特征。这种方法后来被广泛地运用于人格测验项目的编制。

1919年，美国武德沃斯发表了第一个自陈人格量表——个人资料调查表，这开了人格问卷测量之先河。

1920年，罗夏克墨迹测验问世，投射测验由此诞生。

目前，用于人格测量的测验多达数百种，从编制测验的方法和测量的程序来看，人格测量技术的主要种类有自陈问卷法、投射法、评定法、情境法、行为观察法、晤谈法等。

二、人格测量的真实性问题

相对于智力测量来说，人格测量的信度和效度更低一些，这就使得人们有理由提出人格测量的真实性问题。而影响人格测量的真实性的因素除了编写测验项目的技术外，受测者是否真实地回答测验所提出的各种问题也是一个重要因素。运用自陈问卷测量人的人格特征时，通常是要求受测者针对所提的问题在“是”和“否”两个备选选项之间选择一个符合他实际情

况的选项。在这种情况下，受测者虽然清楚他应当选择“是”或“否”，但由于人格结构中的一些特质具有明显的社会评价色彩，受测者为了获得较高的社会评价，或不愿意让其他人了解自己的真实的人格特征，完全可能选择一个与自己实际情况相反的选项，这是其一。其二，有的受测者在某些项目上可能不太清楚哪个选项更符合自己的实际情况，所以在拿不准的情况下，常常随便选择一个选项。其三，有的被试在无意识中就有一种防卫倾向，所以不知不觉地选择了与自己的实际情况不符合的选项。最后，由于目前流行的人格问卷所提供的备选选项太少（通常只是“是”与“否”两种），受测者可能感到任何一个选项都不太符合自己的实际情况。在这种情况下，受测者要么两个选项都选，要么两个选项都不选，或者不加思索任意选择其中的一个。有的测验的编制者（如卡特尔）意识到了这个问题，于是在两个极端的选项之间插入一个折中性选项（如“不一定”、“介于‘是’与‘否’之间”），但实际上，受测者在一个具体的问题上很少有这种不偏不倚的中间情况。

为了防止受测者回答问题时有意识或无意识的防卫性反应，有的问卷插入了一个说谎量表，假如受测者在该量表上的得分过高，则说明受测者没有真实回答，所以其它方面的分数也就不能作为评价他的人格特征的依据。在《明尼苏达多项人格调查表》和《艾森克人格问卷》中就包含了这种说谎量表。但这只能在一定程度上解决测量的真实性问题，假如多数受测者的说谎分数都高，测验就没有多大意义。当然，在实际测量中这种情况很少出现。

防止人格测量不真实的另一个办法是不用自陈问卷法，而改用投射测验。投射测验的一个优点是可以让受测者在不知不觉中将他的无意识心理投射到他对测验项目的反应之中。但目前的投射测验结果很难做到量化、对测验结果的解释是施测者

的主观看法，不同的施测者对同一个测验结果的解释常常不完全相同。因此，假如对测验结果给予不同的解释，那么，尽管测验结果本身是真实的，也难以说明整个测量工作的真实性。

当然，人格测量中存在的上述难以保证真实性的问题并不否定人格测量在一定程度上的科学性，这只是一个进一步改进和完善的问题。在人格测量中尽管存在着一定的难度和复杂性，但经过将近100年的探索和发展，已经初步形成了一套比较科学的人格测量方法和技术，并在实际应用领域发挥着越来越重要的作用。

第二节　自陈量表

自陈人格测量就是根据要测量的人格特质，编制许多有关的问题，要求受测者根据自己的实际情况逐一回答这些问题，然后根据受测者的答案，去衡量受测者在这种人格特质上表现的程度。为完成自陈人格测量而编制的测量工具叫自陈量表或自陈问卷。自陈量表的项目形式一般采用是非式或选择式，它的计分规则比较客观，施测手续比较简便，测量分数容易解释，因此，是人格测量中应用最广的一种测验。

一、自陈量表的编制及其特点

（一）自陈量表的编制

所谓“自陈”就是自我陈述，即让受测者个人提供关于自己人格特征的报告。由于纯粹主观的自我报告对有关的变量难以控制且不易获得客观的数量化的评价，因此自陈法多采用客观测验的形式，也就是测验的编制者预先拟定一系列陈述句或问题，每个陈述句或问题描述一种行为特征。若干个描述行为特征的陈述句或问题组成共同测量一种人格特质的量表。同时，在每一个陈述句或问题之下提供两个或两个以上的选项，供受测者根据自己的实际情况选择。

编制自陈量表的基本假设是只有受测者最了解自己的人格特征。因为个人随时随地都在观察自己的行为，而他人不可能了解自己行为的所有方面。

编制自陈人格量表的第一项任务是确定所要测量的人格特质，并明确给出该特质的操作性定义，然后围绕着该特质选择能够表现该特质的行为情境和反应。具体的编题方法有以下几种：

(1) 是否式：提供一个陈述句或问句，并列出“是”和“否”两种选项，要求受测者选择其中的一个选项。例如：

我喜欢上街游玩。　　是□　否□

你有许多业余爱好吗?　　是□　否□

(2) 二择一式：提供两个意思相反的陈述句（A，B），要求受测者选择其中符合自己实际情况的一个。例如：

A．我常批评那些有权威和有地位的人。　□

B. 在长辈或上级面前，我总是感到胆怯。　□

（3）是否折中式：提供一个陈述句或问句，并列出“是”、“否”和“不一定”（或“介于是与否之间”）三种选项，要求受测者选择其中的一个选项。例如：

我善于控制自己的表情：A. 是的　B. 介于 A 与 C 之间　C. 不是的

（4）文字等级式：提供一个问句，同时列出几个（通常是五个）程度不等的选项，供受测者选择。例如：

你对自己的工作满意吗？

非常满意□　比较满意□　无所谓□　不大满意□　极不满意□

（5）数字等级式：实际上是文字等级式的变式，只不过是将文字式选项改为数字式选项。例如：

你对自己的工作满意吗？

非常满意——————————非常不满意

1　2　3　4　5

如前所述，在运用自陈量表测量人的人格特质时，受测者可能有意无意地选择不符合自己实际情况的选项。为了尽可能防止这种情况的发生，在编写测验项目时，应当注意：①尽可能回避带有明显的社会评价色彩的问题，代之以中性的陈述。例如，我们如要测量人的工作责任感，可以编写诸如“对于生活中的大多数事情，我都要做得妥贴才能放下心来”的陈述，而不要直接了当地编写成“只要是领导安排的工作，我都能保证认真按时地做好”。因为，后者具有明显暗示和社会评价的色彩。②对于量表中必须涉及的个人私生活问题，应当采用适当隐蔽的措辞予以表述。例如，可以编写诸如“事实上，许多人在内心中都怀有一些不可告人的想法”，而不要编写成“你的内心中有一些不可告人的想法吗”。尤其是当涉及个人的性

问题时，所编写的项目更应当做一些技术上的处理，以防止引起受测者的反感而作出虚假的回答。③所提供的选项最好排列成若干个等级，以便受测者选择更接近他实际情况的答案。

（二）自陈量表的特点

（1）自陈量表的题量较大，多数用于测量人格的若干特质。例如，著名的《明尼苏达多项人格调查表》总共有566个是否项目，包含3个效度表和10个临床量表，其中临床量表可以测量人格的10种特质；《卡特尔16种人格因素量表》共有187个项目，用以测量人格结构的16种特质。当然，也有的量表尽管题量较大，但只测人格的一个方面，如《内—外向量表》。

（2）自陈量表通常采用纸笔测验，即将测验项目印在纸上装订成册，另有一张答卷纸，将备选选项印在答卷纸上，被试一边阅读测验项目，一边在答卷上选择适合于自己的选项。这样可以同时测量许多人。近年来，由于计算机的发展和普及，人们为了省去评分和计算上的麻烦，将测验编成计算机程序，受测者直接在机器上作答，计算机根据受测者答题的情况直接打印出测量结果。

（3）自陈量表的计分规则简单而客观，施测手续比较简便，测量分数容易获得解释。因此一般对测验情境和施测者的要求不像智力测验那样严格。

（三）自陈量表的信度和效度

和智力测验一样，标准化的人格量表应当具有测验信度和效度指标的报导，但由于人格特征在行为中的表现远比智力的表现复杂和多样，也由于人格测量中受测者具有较强的防卫性，人格量表的信度和效度比智力测验要低。就目前流行的著

名人格量表而言，信度指标通常采用重测信度和内部一致性信度，其信度系数一般不低于 0.6；而效度指标通常采用理论建构效度，而较少有效标效度的报导，因为在人格测量中较难找到适当而又实用的效标。

二、《明尼苏达多项人格调查表》的使用

（一）《明尼苏达多项人格调查表》简介

《明尼苏达多项人格调查表》（英文简称 MMPI）是由美国明尼苏达大学临床心理学系系主任哈撒韦（S. R. Hathaway）和心理治疗家麦金利（J. C. Mckinley）于 40 年代共同编制的。在编制过程中，他们进行了大量细致的研究工作。首先从大量病史、早期出版的人格量表以及心理医生的笔记中选编了大量的项目，然后对正常人和心理异常被试进行测量，经过重复测量，交叉测量以验证每个分量表的信度和效度。经过临床实践的反复验证和修订，到 1966 年修订版的项目确定为 566 个，其中 16 个项目为重复项目（用于检测受测者反应的一致性）。566 个项目中前 399 个项目分别分配在 13 个分量表中，包括 10 个临床量表和 3 个效度量表；其余的项目则与一些研究量表有关。通常在临床诊断中只使用前 399 个项目。

MMPI 的项目内容范围非常广泛，包括身体各方面的状态（如神经系统、心血管系统、生殖系统等），精神状态以及对家庭、婚姻、宗教、政治、法律、社会等态度。

几十年来，MMPI 一直被广泛应用，翻译成各种版本达 100 余种，应用范围也扩展到诸如心理学、医学、人类学和社会学等领域的研究工作中

在中国，宋维真从1980年开始主持试用修订MMPI，于1989年完成了标准化工作，取得了中国版的信度和效度资料，并制定了中国常模。可用于测量16岁以上具有初中文化程度的中国人。

修订后的项目仍为566个，只是对项目中的个别词句做了适当的改动。10个临床量表的名称及其字母代号见表14.1。

表14.1　MMPI临床量表的名称及其字母代号

序号	量表名称	英文缩写	序号	量表名称	英文缩写
01	疑病	Hs	06	妄想狂	Pa
02	抑郁	D	07	精神衰弱	Pt
03	癔病	Hy	08	精神分裂	Sc
04	精神病态	Pd	09	轻躁狂	Ma
05	男性化—女性化	Mf	10	社会内向	Si

3个效度量表的名称和意义如下：

(1) 说谎量表（L）：分数高表示回答不真实。

(2) 诈病量表（F）：分数高表示诈病或确系严重偏执。

(3) 校正量表（K）：分数高表示一种自卫反应。

此外，在效度量表中，可增加疑问量表（Q），即无法回答的项目数。无法回答的项目数超过一定的标准，则认为此答卷不可靠。

(二)《明尼苏达多项人格调查表》的使用

1. 施测方法

按MMPI题册首页上的指导语进行。在进行测验前，主试应当熟悉全部测验材料（包括调查表的内容、简介、指导

语、信度和效度的资料以及常模资料等)，了解受测者的有关情况（如文化程度、理解能力及身体状况)。测验情境应尽可能安静，没有无关的人在场。如果测验结果用于临床诊断，主试在测验前，一定要让患者了解这个测验的重要性以及对他治疗的好处，以便得到患者的合作。

2. 计分方法

用预先制作的 14 张套板（每个分量表一张，Mf 为两张，男女各一张）进行计分，步骤如下：

(1) 将答卷按受测者性别分开。

(2) 将答卷纸上同一题划有两种答案的题号用彩色笔划去，当作没回答，与“无法回答”的题数相加，作为 Q 原始分数。如果总分超过 30 分，则此答卷无效。

(3) 将每个分量表的套板依次覆盖在答卷纸上对准，数好套板上有多少个圆洞被涂黑，这个数目就是该分量表的原始分数，将此分数登记在此量表的原始分数栏内。

(4) 在疑病 (Hs)、精神病态 (Pd)、精神衰弱 (Pt)、精神分裂 (Sc) 和轻躁狂 (Ma) 5 个分量表的原始分数上加 K 分，方法是 Hs + 0.5K，Pd + 0.4K，Pt + 1.0K，Sc + 1.0k，Ma + 0.2K（注意：字母所表示的分数均为原始分数)。不过，对于中国被试，加或不加 K 分，对测量的总结果没有什么明显影响，可以不加 K 分。

(5) 将各分量表的原始分数转记在剖面图的原始分数栏内。

3. 原始分数的转换

MMPI 的常模采用 T 分数。在分数的转换过程中，先将受测者在各分量表上的原始分数根据常模表，转化成相应的 T 分数，登记在剖面图的 T 分数栏内；然后在剖面图上找到各分量表 T 分数的点，将各点相连，就成为一条表示受测者人

格特征的曲线图。

4. 测量结果的解释

对各分量表的T分数可参照MMPI说明书中对各分量表分数提高的意义的文字描述予以解释。这里需要强调的是，说明书中所列举的人格特点只是一类人的共同的典型的特点，在具体地解释一个人的分数时应当持慎重和灵活的态度。这一原则同样适用于其他人格量表。

三、《卡特尔16种人格因素量表》的使用

（一）《卡特尔16种人格因素量表》简介

《卡特尔16种人格因素量表》（简称16PF）是由美国伊利诺州立大学教授雷蒙德·B·卡特尔（Raymond B. Cattell）经过几十年的系统观察，科学实验以及因素分析统计后逐渐形成的。这一量表能在约45分钟的时间内测量出16种主要的人格特质。初中以上文化程度的人均可接受本量表的测试。

16PF在国际上广泛流行，现已译成法、意、德、日、中等多种文字，被许多国家修订。16PF中的16种人格因素是各自独立的，每种因素与其他因素的相关度较小。借助于本量表，受测者不仅可以对自己在16个因素上的人格特点获得了解，而且根据卡特尔制定的人格因素组合公式可以对自己的整体人格做出评价。

16PF英文版有A、B两套等值的测题，每套187个项目，分配在16个因素中。每个因素所包含的项目数不等，少则13个，多则26个。每个项目有a、b、c 3个选项（如a：是的；b：不一定；c：不是的）、受测者根据自己的情况选择一个合

适的选项。

16PF中国版的修订工作是在辽宁省修订本的基础上由戴忠恒与祝蓓里主持完成的，取得了全国范围内的信度和效度资料，制定了中国成人（男、女）常模、中国大学生（男、女）常模、中国中学生（男、女）常模、中国产业工人常模、中国专业技术人员常模、中国干部常模以及上海市的各种常模。

16PF所测量的人格因素的名称及其字母代号见表14.2。

表14.2　16种人格因素的名称及其字母代号

代号	因素名称	代号	因素名称	代号	因素名称	代号	因素名称
A	乐群性	F	兴奋性	L	怀疑性	Q_1	实验性
B	聪慧性	G	有恒性	M	幻想性	Q_2	独立性
C	稳定性	H	敢为性	N	世故性	Q_3	自律性
E	恃强性	I	敏感性	O	忧虑性	Q_4	紧张性

（二）《卡特尔16种人格因素量表》的使用

1. 施测方法

16PF属团体测验。施测时，先给每个受测者发一张答卷纸，填上受测者的姓名、性别、年龄、职业、测验日期等。然后发给测题，翻到测题的说明部分，让受测者边看边听主试朗读其中的指导语，并在主测的指导下完成答卷纸上方的4个例题，待受测者掌握答题方法后，即让受测者自己完成正式测验。对施测情境的要求与MMPI相同。

2. 计分方法

每个项目有a、b、c 3个选项，根据受测者对每一项目的回答，分别记为0，1，2分或2，1，0分。实际操作时，要用预先制作的两张有机玻璃计分套板，每张套板记8个因素的分

数。方法是：将套板套在答卷纸上，分别计算出每一因素上的原始分数，将此分数登记在剖面图左侧的原始分数栏内。

3. 原始分数的转换

16PF 的常模采用标准 10 分制。根据受测者的文化程度或职业种类将受测者各因素的原始分数对照常模表分别转化成标准分数，并登记在剖面图左侧的标准分数栏内。然后在剖面图上找到各因素的标准分数点，将各点相连，即成为一条表示受测者人格特征的曲线图。

4. 测量结果的解释

根据剖面图上对各因素高分特征和低分特征的描述，可以大体解释受测者在 16PF 上的主要特点。但如要作进一步的解释，则需参照《16PF 手册》中的文字描述。

16PF 不仅能够对受测者在 16 种人格因素上的主要特征进行分析性描述，而且能够根据实验统计结果所得的 4 个公式对他在次级人格因素上的特征（分别用于诊断受测者的适应性、外向性、情绪性和果断性）进行综合描述。同时，可以利用另外 4 个公式预测受测者在某些特殊情境中的行为特征（即心理健康水平、专业成就的可能性、创造潜力、对新环境的适应能力），尤其适用于升学、就业及生活问题的指导。

三、《艾森克人格问卷》的使用

（一）《艾森克人格问卷》简介

《艾森克人格问卷》（英文简称 EPQ），由英国心理学家汉斯·艾森克（H.J.Eysenck）和其夫人于 1975 年在先前几个人格调查表的基础上编制。它的理论基础是艾森克所提出的人

格三维度理论。艾森克认为，虽然人格在行为上的表现形式是多样的，但真正支配人行为的人格结构却是由少数几个人格维度构成的。艾森克经过长期的实验研究和临床观察，提出精神质、外倾性和神经质是人格的三个基本维度。这里，人格维度代表着一个连续体，每个人都或多或少地具有这三个维度上的特征，但不同的个人在这三个维度上的表现程度是不同的。因此，通过测量可以在这些维度上找到受测者的特定位置。根据这种观点编制的 EPQ 由 4 个分量表构成（P、E、N 和 L），用于测量受测者在精神质（P）、外倾性（E）和神经质（N）三个人格维度上的特征。L 是说谎量表，用于识别受测者回答问题时的诚实程度。该问卷分儿童和成人两种，儿童问卷共有 97 个项目，适用于 7～15 岁的受测者，成人问卷共有 101 个项目，适用于 16 岁以上的受测者。

EPQ 中国版由龚耀先教授主持修订。修订后的儿童问卷和成人问卷各由 88 个项目组成。每个项目都有“是”和“否”（在儿童问卷中是“是”和“不是”）两个选项，供受测者选择。他们通过标准化工作，取得了全国范围内的信度和效度资料，制定了中国儿童（男、女）和成人（男、女）常模。

（二）《艾森克人格问卷》的使用

1. 施测方法

EPQ 属于团体测验。施测时，先给每个受测者发一张答卷纸，填上受测者的姓名、性别、年龄、测验日期、职业、文化等。然后发给测题，翻到测题的说明部分，让受测者边看边听主试朗读其中的指导语，待受测者掌握答题方法后，即让受测者自己完成正式测验。对施测情境的要求与 MMPI 以及 16PF 相同。

2. 计分方法

EPQ 记分的依据是记分键（见该问卷的《手册》）。记分键中的数字是项目号，项目号前无“-”号的表示该项目若受测者圈“是”记1分，圈“否”（或“不是”）记0分；项目号前有“-”号的表示该项目若受测者圈“否”（或“不是”）记1分，圈“是”记0分。按 P、E、N、L 4个分量表分别记分，然后算出各分量表的总分（原始分数）。

3. 原始分数的转换

EPQ 的常模采用 T 分数。根据受测者的性别和年龄将受测者各分量表的原始分数对照常模表分别转化成 T 分数，然后在剖面图上找到各维度的 T 分数点，将各点相连，即成为一条表示受测者人格特征的曲线图。

4. 测量结果的解释

对精神质（P）、外倾性（E）和神经质（N）三个人格维度上受测者的 T 分数的解释可参照《手册》中对高分特征和低分特征的文字描述。

此外，艾森克还将外倾性（E）和神经质（N）两个维度作了垂直交叉分析，这样就可以得到4种典型的人格类型，它们的名称及其主要特征如下：

(1) 外向稳定型：善领导，无忧虑，活泼，悠闲，易共鸣，健谈，开朗，善交际。

(2) 外向易变型：主动，乐观，冲动，易变，易激动，好斗，不安定，易怒。

(3) 内向易变型：文静，不善交际，缄默，悲观，严肃，刻板，焦虑，忧郁。

(4) 内向稳定型：镇静，性情平和，可信赖，有节制，平静，深思，谨慎，被动。

除以上4种典型的人格类型外，还有多种变型．根据受测

者 E 和 N 的 T 分数可以在剖面图上找到相应的交点。

四、《学生性格量表（11～18 岁）》的使用

（一）学生性格量表（11～18 岁）简介

《学生性格量表（11～18 岁）》（英文简称 SPS），由云南师范大学沙毓英、张锋等人共同编制，适用于测量中国 11～18 岁中小学生的性格（亦即人格）特征。该量表于 1992 年通过专家委员会的鉴定，并于 1995 年完成了云南省常模的编制工作。

SPS 的理论基础是编制者提出的中国人性格层次结构理论。他们认为，性格是多层次多因素的整合结构，在横向上可分解为彼此相联系的多种特质因素，在纵向上可分解为由抽象到具体的多种层次。性格的最高层是整合层，也就是性格本身。第二层是集质层，分解出性格的 5 个亚结构，即：①生活旨趣：指个人对生活目标和生活价值的追求；②认知风格：指个人认知事物、思考问题的方式；③情绪特征：指个人的情绪反应特点；④意志品质：指个人的意志行为特点；⑤态度倾向：指个人对集体、他人和自我的对待方式。第三层是特质层，每个亚结构又具体分解为若干种性格特质。生活旨趣被分解为实惠性、知识性、支配性和奉献性四种特质；认知风格被分解为客观性、全面性、独立性、简略性和敏捷性五种特质；情绪特征被分解为激活性、强烈性和持续性三种特质；意志品质被分解为自觉性、自制性、坚持性、果断性和敢为性五种特质；态度倾向被分解为责任感、荣誉感、进取性、利他性、真诚性、攻击性和外倾性七种特质。总共 24 种性格特质。第四

层是行为层，指的是典型情境中的典型行为反应，是最能反映特质的可观察的有代表性的行为。性格特质表现在个人对不同情境的类似反应之中，因此通过编制量表测量个人在典型情境中的典型行为，就有可能推断出个人的性格特质。

根据上述理论观点编制的《学生性格量表（11～18岁）》共有168个项目。每种特质包括7个项目，为一个分量表，共24个分量表。每个项目均是对一个行为情境的文字描述，每个项目之下是对4种常见的典型行为反应的文字描述，供受测者选择。例如：

和一个人初次打交道时，你会：

（1）觉得不自在，无话可说。

（2）觉得拘束，找不到多少话说。

（3）很快就能和他熟悉起来。

（4）很快就能和他成为好朋友。

SPS在张锋的主持下现已获得了在云南省范围内的信度和效度资料，并分别制定了云南省城市汉族、农村汉族和农村少数民族的小学生（男、女）常模、初中生（男、女）常模和高中生（男、女）常模。

通过对小学五年级（N = 62，男女各半，间隔两个月）、初中二年级（N = 66，男女各半，间隔两周）和高中二年级（N = 86，男女各半，间隔两周）实施重测，获得各特质的重测相关系数如表14.3。

表14.3 SPS各特质的重测相关系数

特质名称	小学	初中	高中	特质名称	小学	初中	高中
实惠性	0.618	0.517	0.712	自觉性	0.579	0.738	0.569
知识性	0.521	0.755	0.567	自制性	0.564	0.685	0.624

支配性	0.567	0.555	0.617	坚持性	0.584	0.580	0.657
奉献性	0.557	0.582	0.668	果断性	0.513	0.679	0.615
客观性	0.560	0.576	0.524	敢为性	0.508	0.564	0.601
全面性	0.675	0.583	0.587	责任感	0.576	0.542	0.783
独立性	0.585	0.553	0.532	荣誉感	0.631	0.538	0.818
简略性	0.593	0.598	0.542	进取性	0.559	0.638	0.605
敏捷性	0.598	0.588	0.554	利他性	0.596	0.617	0.547
激活性	0.558	0.627	0.616	真诚性	0.618	0.604	0.567
强烈性	0.534	0.609	0.697	攻击性	0.585	0.681	0.557
持续性	0.626	0.656	0.741	外倾性	0.574	0.711	0.724

通过对小学五、六年级（N＝124，男女各半），初中一、二、三年级（N＝238，男女各半）和高中一、二、三年级（N＝260，男女各半）施测，获得各特质内部一致性信度系数如表 14.4。

表 14.4　SPS 各特质的内部一致性信度系数

特质名称	小学	初中	高中	特质名称	小学	初中	高中
实惠性	0.789	0.804	0.788	自觉性	0.819	0.830	0.802
知识性	0.720	0.830	0.774	自制性	0.811	0.840	0.820
支配性	0.648	0.778	0.759	坚持性	0.840	0.785	0.766
奉献性	0.707	0.850	0.802	果断性	0.860	0.830	0.819
客观性	0.637	0.766	0.751	敢为性	0.783	0.779	0.802
全面性	0.628	0.771	0.744	责任感	0.795	0.829	0.826
独立性	0.746	0.769	0.810	荣誉感	0.786	0.850	0.857
简略性	0.604	0.737	0.741	进取性	0.899	0.887	0.832
敏捷性	0.759	0.855	0.824	利他性	0.808	0.837	0.808

激活性	0.768	0.786	0.783	真诚性	0.804	0.840	0.811
强烈性	0.890	0.807	0.810	攻击性	0.846	0.786	0.779
持续性	0.820	0.842	0.833	外倾性	0.893	0.835	0.837

SPS的效度指标采用建构效度，分别对小学五、六年级（N=124，男女各半），初中一、二、三年级（N=238，男女各半）和高中一、二、三年级（N=260，男女各半）施测，获得三个学段各特质之间的相关系数。结果表明，除少数特质之间具有中等程度的相关外，大多数特质之间的相关较低或接近零相关。因此，多数因素是相对独立的性格特质。

（二）《学生性格量表（11～18）》的使用

1. 施测方法

SPS属于团体测验。施测时，先给每个受测者发一张答卷纸，并填写有关信息。然后发给测题，翻到题册第1页的说明部分，让受测者边看边听主试朗读其中的指导语，并在主试指导下完成3个例题，待受测者掌握答题方法后，即让受测者自己完成正式测验。对施测情境没有特殊要求，一般在学生自己的教室里就可以了。

2. 计分方法

SPS的计分规则是有些分量表的项目按受测者所选的答案1、2、3、4分别记0、1、2、3分；有些分量表的项目按受测者所选的答案1、2、3、4分别记3、2、1、0分。将受测者在每个分量表项目上的得分加起来，登记在答卷纸的“备注”一栏，即为其特质总分（原始分数），最低分0分，最高分21分。

3. 原始分数的转换

SPS的常模采用T分数，先将受测者的原始分数转登在

剖面图的原始分数栏内，然后根据受测者的民族、性别和学段将受测者各分量表的原始分数对照常模表分别转化成 T 分数，在剖面图上找到各特质的 T 分数点，将各点相连，即为表示受测者性格特征的曲线图。

4. 测量结果的解释

对受测者在 24 种性格特质上的 T 分数的解释可参照《手册》中对高分特征和低分特征的文字描述。需要说明的是，SPS 目前的常模资料来自云南省城乡，其它地区在使用时应当充分考虑它的局限性。而作为研究工具，可以借此进行不同地区间的群体比较。

第三节　投射测验

一、投射测验及其理论基础

（一）投射测验的性质及其特点

投射（projection）是指个人对客体特征的想象式解释，在这种解释中，个人具有将自己身上发生的心理过程无意识地附着在客体身上的倾向。换句话说，投射是个人把自己的思想、态度、愿望、情绪、性格等心理特征无意识地反应在对事物的解释之中的心理倾向。由于心理投射的作用，人们常常把无生命的事物看成是有生命的事物，把无意义的现象解释成有意义的现象。在这种情况下，个人对客体特征的投射性解释所反映的不是客体本身的性质，而是解释者自己的心理特征。因此，

运用投射技术测量个人对特定事物的主观解释，就有可能获得对受测者人格特征的认识。

投射技术作为一个心理测量术语，是1938年由主题统觉测验的编制者莫瑞最早提出的，但投射测验作为一种心理测量技术早在1921年之前就已有人开始探索并实际应用了。1921年，罗夏克发表了他编制的墨迹测验，当时未引起人们的重视。1938年，弗兰克（L.K.Frank）明确阐述了投射技术的内涵及其重要性，他认为投射技术能够唤醒被试内心世界或人格特征的不同表现形式，从而在对测验项目的反应中投射出被试内在的需要和愿望。

投射技术的基本方式是向受测者提供预先编制的一些未经组织的、意义模糊的标准化刺激情境，让受测者在不受任何限制的情况下，自由地对刺激情境作出他的反应，然后通过分析受测者的反应，推断受测者的人格特征。按照这种技术编制的最为著名的人格测验是罗夏克墨迹测验和莫瑞主题统觉测验。

投射测验的特点是：①测验材料没有明确的结构和确切的意义，这就为受测者提供了针对测验材料进行广阔自由联想的机会和空间；②受测者对测验材料的反应不受限制，可以根据自己对测验材料的理解作任何想象式解释，因此受测者对测验材料的解释在很大程度上不是决定于测验材料的性质，而是决定于受测者的人格特征和当时的心理状态；③测验的目的具有明显的隐蔽性，受测者事先并不知道施测者对他的反应作何心理学的解释，这就在很大程度上避免了受测者的伪装和防卫，使测验的结果更能反映受测者真实的人格特征；④对测验结果的解释重在对受测者的人格特征获得整体性的了解，而不是对某个或某些单个人格特质的关注；⑤投射测验的内容多为无明确意义的图片，在测验时不受语言文字的限制，所以，被广泛地应用于人格的跨文化研究；⑥相对于自陈量表，投射测验的

最大局限是计分上的困难，这使得研究者对测验结果难以进行确定的定量分析。

（二）投射测验的理论基础

投射测验重在探讨人的无意识心理特征，对受测者在测验上反应的解释就不可避免地受到精神分析理论的影响。按照精神分析理论的无意识观点，个人无法单凭自己的意识功能了解到自己的人格特征，因此，运用自陈问卷法不可能测量到受测者的真实的人格特征。如果我们以某种无确定意义的刺激情境作为引导，受测者就会在不知不觉中将自己无意识结构中的愿望、要求、动机、心理冲突等特征投射在对刺激情境的解释中。

从上述理论出发，投射测验假定：①人们对外部事物的解释性反应都是有其心理原因的，同时也是可以给予说明和预测的；②人们对外部刺激的反应虽然决定于所呈现的刺激的特征，但反应者过去形成的人格特征、他当时的心理状态以及他对未来的期望等心理因素也会渗透在他对刺激的反应过程及其结果之中；③正因为个人的人格会无意识地渗透在他对刺激情境的解释性反应之中，所以，通过向受测者提供一些意义模糊的刺激情境，让受测者对这种情境做出自己的解释，然后通过分析他解释的内容，就有可能获得对受测者自身的人格特征的认识。

（三）投射测验的信度和效度

虽然投射测验在国外被广泛地应用于对人格特征的评价过程中，尤其是本世纪 40 年代至 60 年代的临床心理学工作者更是把它视为临床诊断中不可缺少的工具。但是，对投射测验的批评却一直没有停止过　除了谈到操作此种测验的技术极度复

杂，难以掌握，难以获得数量化的常模资料外，最为严重的批评莫过于对投射测验的信度和效度持质疑态度。以罗夏克墨迹测验为例，虽然有研究资料认为它的信度和效度是不错的，但是更多的研究却证明它的信度和效度都很低。导致这些相互矛盾的研究结果的一个原因是投射测验本身的性质决定了难以获得确切的信度和效度资料，也难以在不同的测验结果之间进行有效的比较。所以，目前投射测验的应用在走下坡路。在我国，除了龚耀先对罗夏克墨迹测验在小范围内做过试用外，对投射测验的研究和应用工作尚未展开。本章以下部分所介绍的罗夏克墨迹测验和莫瑞主题统觉测验只供有兴趣的读者了解，而不能作为开展这项工作的技术依据。

二、罗夏克墨迹测验简介

（一）罗夏克墨迹测验的形成

罗夏克墨迹测验是由瑞士精神病学家罗夏克（H. Rorschach）经过长期的试验和比较研究后创制的一种投射测验。他从1910年开始用画片来研究精神障碍对病人知觉过程的影响，后来改用墨迹图。在最初制作墨迹图时，他先在一张纸的中央滴一堆墨汁，然后将纸对折，并用力挤压，从而形成两边对称但每次形状不一的图形。罗夏克用大量这样的墨迹图片对各种精神病人进行试验，发现不同类型的精神病人对墨迹图片的反应不同，然后再和低能者、正常人、艺术家的反应作比较，最后选定其中的10张墨迹图片作为测验材料，并确定了记分方法和解释测验结果的原则，于1921年正式发表。10张墨迹图卡中，有5张是黑白的，有3张是彩色的，另有2

张是除黑色外，还带有鲜明的红色。

（二）实施罗夏克墨迹测验的基本程序

实施罗夏克墨迹测验是一项极度复杂的工作，只有那些经过专门的培训，并具有丰富临床经验的人员才能使用。这里介绍的只是其中最基本的实施程序。

1．指导语

在施测之前，主试应当向受测者提供一个简短的指导语：要给你看的图卡上印刷着偶然形成的墨迹图象；请你将看到图卡时所联想到的东西，不论什么，都自由地、原封不动地说出来；回答无所谓正确与不正确，所以，请你看到什么就说什么。

2．施测

施测过程分 4 个阶段：

①自由反应阶段：让受测者对所看到的墨迹图的内容进行自由联想，主试原原本本地记录受测者的所有言语反应。在这一阶段，主试与受测者之间一般不应交谈。

②提问阶段：在这一阶段，主试为了对受测者的反应进行记号化，有针对地向受测者提出问题。

③类比阶段：当利用经过提问获得的资料仍不能搞清记号化的问题时，可在类比阶段作进一步的商讨。

④极限测验阶段：在这一阶段，主试对受测者的反应产生疑问时，进行进一步确认。

3．记号化

记号化是指对受测者的测验资料进行分类，将具有相似特性的反应归类，并给予同样的记号。记号化包括 4 个方面：

①区位记号：这是根据受测者对墨迹图反应的范围进行的分类，有 5 种类别：整体反应（W）、普通局部反应（D）、细

微局部反应（d）、特殊局部反应（Dd）和空白反应（S）。

②决定因子记号：这是根据受测者对墨迹图反应时的依据所作的分类，有4个方面：形状反应（F）、运动反应（M）、浓淡反应（K）和色彩反应（C）。

③内容记号：这是根据受测者对墨迹图所作的反应的内容进行的分类，主要有以下典型的反应内容：人（H）、动物（A）、解剖（At）、性（Sex）、自然（Na）、物体（Obj）等等。

④独创记号：这是根据受测者对墨迹图反应的独特性所作的分类，有普通反应（P）和独创反应（O）两种情况。

4. 测验结果的解释

根据上述记号化的结果，在决定因子的心理图像上标上每个因子的反应次数，将各点相联，即是受测者的人格图像。然后结合反应区位、反应内容、反应的独创性，以及它们之间的数量关系，根据测验手册中的描述，解释受测者的人格特征。

一般来说，W分高，表示具有高度的综合能力，但过高也表明缺乏精细分析的能力；M分高，表示具有想象力和移情倾向；C分高，表示性格外向，情绪不稳定；A分高，且反应资料呈无组织的状态时，表示智力低下，思维刻板；F分高，表示具体良好的自我控制能力和情绪活动的和谐；K分高，可能预示着不安的情绪，等等。在对各记号项目进行解释时，应注意对各种分数作综合性的解释，不可凭任何单一的分数来判断一个人的人格是否正常。只有这样，才能体现投射测验的初衷。

三、主题统觉测验简介

(一) 主题统觉测验的形成

主题统觉测验(Thematic Apperception Test, 简称 TAT)是另一种与罗夏克墨迹测验齐名的人格投射测验,它是由美国哈佛大学的心理学家莫瑞和摩尔根于1935年创制的,此后经过三次修订。TAT是一种窥探受测者的主要需要、动机、情绪、情操和人格特征的方法。它的基本原理是向受测者呈现一系列意义相对模糊的图卡,并鼓励他按照图卡不加思索地编述故事。编制这种测验的基本假设是:①人们在解释一种模糊的情境时,总是倾向于将这种解释与自己过去的经历和目前的愿望相一致;②在面对测验卡讲述故事时,受测者同样利用了他们过去的经历,并在所编造的故事中表达了他们的感情和需要,而不论他们是否意识到这种倾向。

现在使用的TAT是经莫瑞修订过的第三版。第三版的全套测验包括30张黑白图卡和1张空白卡,图卡的内容有的为人物,有的为景物。就测验内容而言,TAT比之罗夏克墨迹测验的组织和意义要明确,但TAT同罗夏克墨迹测验一样,对受测者的反应不加任何限制,任其针对图卡凭自由想象去编造故事。30张图卡分为四组,分别是成年男性组(M)、成年女性组(F)、儿童男性组(B)和儿童女性组(G)。其中有的图卡适用于所有的受测者(只用数字表示顺序号),有的图卡只适用于特定年龄及特定性别的受测者(分别用数字后面的字母标明)。适用于各组受测者的图卡均为19张,外加1张空白卡,共20张图卡。

（二）实施主题统觉测验的基本程序

1. 测验环境与指导语

测验应当在友好的气氛当中进行，主试对于受测者的反应应当持有鼓励和赞许的态度；测验环境布置应当具有一定的情调，并能激发人的想象力和创造性。一般的指导语是：这是一个想象力的测验，是测验你的智力的一种形式。我将让你看一些图片，每张都让你看一会儿。你的任务是对每张图片尽你所能，编一个带有戏剧性的故事，说明是什么因素导致了图片上的情景，当前在发生什么事情，图片上的人正在想什么，结果会怎么样。你可以用5分钟讲一个故事。

2. 施测

在实施TAT时，每个组的受测者都要完成两个系列的测验。第1～10号图卡为第一系列，第11～20号图卡为第二系列。其中第二系列图卡的情境更加抽象，也更加奇特。完成每个系列的测验任务需要1小时的时间，两个系列之间至少要间隔一天。在测验过程中，主试要记录受测者所说的内容，如果笔记有困难，可以利用录音机录音，前提是不能让受测者发觉。

3. 评分

TAT的评分分两部分：一是在每一种需要变量和情绪变量上的分数，评分规则是根据每一种需要或情绪的强度在1～5之间记分；二是在每一种压力变量上的分数，评分规则是根据每一种压力的强度在1～5之间记分。最后在每一变量上都得到两个分数，一是总体平均分（AV），二是分数的分布（R）。

被评定的主要的需要变量、情绪变量有：恭顺、成就、攻击、自责、关怀、顺从、性、受保护、进取、归属、自主、矛

盾、情绪变化、沮丧、焦虑、怀疑等；被评定的主要的压力变量有：归属、攻击、支配、关怀、拒绝、身体危险等。而评定这些变量的分数的依据是受测者在所编的故事中对主人公的行为、需要、动机、情感和主人公所处的环境的描述，以及整个故事所反应出的主题的性质。

4. 测验结果的解释

解释 TAT 分数有两个基本假设：第一个假设是主人公的归因（需要、情绪状态和情感）代表着受测者人格的倾向性。这种倾向性是受测者的过去和他所预期的将来，即：①他已做过的事；②他想去做的事；③他未意识到的一些基本的人格力量；④他当时所体验的情绪和情感；⑤他对将来行为的预测。

第二个假设是受测者所统觉的环境压力也代表着过去、现在和将来，即：①他真正遇到过的情境；②他出于愿望或恐惧而想象到的情境；③他正在统觉的情境；④他期望遇到的或害怕遇到的情境。

主试应当根据上述两个基本假设参照手册中对各种需要、情绪及压力变量的基本描述去解释受测者投射在所编的故事中的人格状态和特征。同时要特别在需要、情绪的力量和压力的力量之间进行强度上的比较，并分析它们之间的相互作用所导致的结果。

练习与思考

1. 试综合比较自陈量表和投射测验的异同。

2*. 查阅有关人格心理学著作，讨论人格的理论研究对发展人格测量技术的作用。

3*. 查阅有关文献，分析中国人格测量研究的现状和特点，并论述你对开展人格测量与研究工作的看法。

第十五章　其他心理与教育测验

本章提要：

●焦虑及其测验方法

●兴趣及职业兴趣测验

●态度和品德的常用测量方法

第一节　焦虑测验

一、焦虑概述

广义地说，焦虑是一种情绪。从强度上看，它涉及到轻重不等但性质相同的相互过渡的一系列情绪，最轻的是不安和担心，其他是害怕和惊慌，最重的是极端恐怖；从快感度上看，它是一种负性情绪，给人的体验是不愉快的；从复杂度上看，它是一种复合情绪，包含有悲哀、恐惧、愤怒等成分。

如果程度恰当，并主要针对某些特定的情境，焦虑是一种正常的、具有适应意义的负性情绪状态，这种体验的作用是向个体报告对外界情境的不适宜，驱使个体采取应付策略或行动，去改变自身的处境；如果焦虑成为自由浮动的、泛化的、或程度过强，便会成为一种异常状态。焦虑可以是一时的情绪状态，也可内化为稳定的个体情绪特质，这样的人性格十分脆弱，而严重的、持续的焦虑，有可能形成病态人格。

焦虑的表现主要在3个方面，一是行为上的表现，如说话唐突、语无伦次、皮肤变红、脸面痉挛、笨手笨脚、结结巴巴、思绪不清等；二是生理上的表现，如肌肉僵硬、全身或局部疼痛、呼吸不畅、心律不齐、寒颤、出汗、排尿过频、食欲减退、失眠、腹泻拉痢等；三是心理上的体验，如烦躁、不安、恐惧、担心等。

对焦虑研究比较早的要属精神分析学派的创始人弗洛伊德

(S. Freud)，他不仅描述了焦虑的表现，还试图解释焦虑的形成过程。弗洛伊德按照他的人格结构理论，认为焦虑是被压抑的性紧张即里比多（Libido）的释放，由于里比多的能量不允许正常释放，一旦累积就要求自动释放，这便会形成焦虑或焦虑症状①。

对焦虑研究起推动作用的是毛瑞（Mowrer）②，他于1939年在《心理学评论》上发表文章，用刺激——反应理论来解释弗洛伊德的"焦虑"。他把焦虑和恐惧看成同义语，并把恐惧定义为产生痛苦反应的条件刺激，由于恐惧反应是一种强烈的不愉快的体验，因而可看成是激发行为和强迫新习惯获得的内驱力。毛瑞把精神分析与学习理论相结合，使焦虑的研究在心理学实验室研究中变得多起来，这些研究表明，恐惧减少有利于激发学习各种条件反射的动机。

受以上观念的启发，为了研究焦虑对人的学习影响，泰勒（Taylor）从MMPI中挑选一些项目，编制了显性焦虑量表（Manifest Anxiety Scale，简称MAS），以研究个体的焦虑水平（动机差异）对瞬眼条件反射学习的影响，即把MAS测得的焦虑得分看成是一种内驱力强度。

与泰勒同时代的研究者还有曼德勒（G. Mandler）和萨拉森（S. Sarason），他们于1952年发表了《测验焦虑问卷》（Test Anxiety Questionnaire，简称TAQ）。曼德勒经过多年研究，还提出了自己的焦虑理论。他认为焦虑是在人处于无助感之中时产生的，"阻断"（interruption）是他观点的核心，任何

① Spielberger, C. D. (1975) Axiety: State - tralt - process. In C. D. Spielberger & I. G. Sarason (Eds), Stress and anxiety, Vo.1 New York: Hemisphere. p116 - 141.

② Spielberger, C. D. (1975) Axiety: State - tralt - process. In C. D. Spielberger & I. G. Sarason (Eds), Stress and anxiety, Vo. 1 New York: Hemisphere. p116 ~ 141.

情景在阻断了或威胁着要阻断已组成的反应系列并且又不能提供任何可替代的反应时，就将引起焦虑。

卡特尔（R.B.Cattell）和同事塞欧（I.H.Scheir）（1961）在对人格的研究方面，亦十分重视焦虑的研究。首先他们发现正常人与神经症患者在焦虑上有差别（当然不是唯一的差别）；其次，他们提出了两种焦虑形式，即特质焦虑（trait anxiety）和状态焦虑（state anxiety），但他们对二者关系的认识尚不是十分清楚。

对焦虑状态和特质研究比较深入的要数施皮尔伯格(Spielberger)，他提出了焦虑的特质——状态理论。所谓焦虑状态是指由紧张、担忧、神经过敏和忧虑所引起的主观感受和由主性神经系统的唤醒（或激发）所引起的生理反应，它发生于某一时刻，有一定的强度水平，但持续时间较短。特质焦虑则是一种比较稳定的人格特质，它存在着个体差异，是一种习得的行为倾向。特质焦虑既可以在过去已有过的焦虑状态的频率和强度上反映出来，也可以在将要经历的未来事件的焦虑状态上反映出来。一般而言，焦虑特征越明显，个体在受到威胁的情景中经历焦虑状态的可能性越大。

二、焦虑测验

焦虑各种各样，因此对焦虑的测量也种类繁多。焦虑分类有以下几种：按焦虑的跨情境程度分，有一般焦虑（即特质焦虑）和特定焦虑（如考试焦虑、怯场、社会交往焦虑等）；按意识程度分，有显性焦虑（意识到）和潜伏焦虑（意识不到）；按其效果分，有积极的焦虑和有害的焦虑。由于焦虑测验较

多，这里仅简要介绍几种。

（一）显性焦虑量表（MAS）

MAS是泰勒[1]按理论推理而建构的量表，她当时编制这个量表主要是为了研究焦虑对学习的动机或驱力作用。泰勒根据卡默龙（N.A.Cameron）（1947）关于慢性焦虑反应所描述的显性焦虑概念，让5位专家（临床工作人员）根据卡默龙的定义来评价MMPI中的项目．如果某项目被判断能反映焦虑的程度达到65%以上，就把其看成能反映显性焦虑，按照这个标准，她从MMPI中获得了65个项目，另外她还加入135条缓冲项目，这些项目也经过了5位专家的评定，一致显示它们不能反映显性焦虑，这便构成了最初的MAS。随后，此量表又进行了多次修订，最后从65个项目中选取了50个项目，把缓冲项目增加到225个，并采取了MMPI中的L、K和F量表中的项目。关于该测验的重测信度，在间隔3周时皮尔逊相关系数为0.89，间隔为5个月的相关系数为0.82，间隔为9~17个月的相关系数为0.81。

为了使MAS适合于大学文化程度以下的人，泰勒等又简化了焦虑项目中的某些难于理解的措词和句子，修订后有28个焦虑项目，而且有两个替代本（复本），这些项目以“是否符合自己的状况”而回答。

（二）《状态——特质焦虑量表》（简称STAI）

《状态——特质焦虑量表》（State——Trait Anxiety Inventory）是由施皮尔伯格等人根据他的理论编制的，首版STAI（X）于1970年问世，作者于1979年对STAI（X）进行修订，

---

① 陈仲庚等：《人格心理学》，沈阳，辽宁人民出版社1987年出版，第112~113页，第384~389页。

1980年修订版称为STAI（Y）。

该问卷的内容包括两个部分，一是状态焦虑，即评定人们“现在”或最近一个特定时间内的感受或人们将要遇到特别情景时的感受；二是特征焦虑，即评定人们通常情况下的情绪体验。

STAI不仅有适合初中、高中、大学生和成年人的状态与特质焦虑量表，另外还开发了适合于小学生的儿童状态——特征焦虑量表（简称STAI C）。状态——特征量表目前已被译成30多种语言，在全世界广泛使用。

叶仁敏（1990）[①] 将STAI（Y）和STAI C在中国进行了修订。该量表是自陈形式，适用于个别或团体施测，无时间限制。状态焦虑量表与特征焦虑量表是分开编制的，各有20个题目，分别做每个测验约需6～10分钟，一起做，共需10～20分钟。如果两个测验都做，最好是先做状态焦虑测验，后做特征焦虑量测验，因为状态焦虑对施测情境敏感，如先测特征焦虑，会形成一定的测试气氛，使状态焦虑测验的结果受影响，而有研究表明，特征焦虑量表几乎不受所给情境的干扰。

STAI的项目计分是4级计分，对焦虑的表述有正反两个方面，对反向表述，计分要反转，这是在计分时应注意的。

STAI按状态焦虑和特征焦虑分别为大学生、中学生以及在职成人的不同性别群体建立了常模，30天间隔的重测信度情况如表15.1。

① 叶仁敏：《状态——特质焦虑量表（Y版）指导手册》，1990年出版

表 15.1　STAI 重测信度表

	中学生				大学生			
	状态焦虑		特征焦虑		状态焦虑		特征焦虑	
	男	女	男	女	男	女	男	女
人数	49	47	49	47	52	36	52	36
相关值	.46	.47	.67	.76	.52	.61	.69	.78

该测验与相关量表的相关在 0.41～0.85 之间，表明有较高的同时效度。

（三）测验焦虑量表

测验焦虑（Test Anxiety）也译为考试焦虑，关于测验焦虑测量研究比较早的要属曼德勒和萨拉森，他们于 1952 年发表了测验焦虑问卷（Test Anxiety Questionnaire，简称 TAQ)。近年来，除 TAQ 外，萨拉森还编制了测验焦虑量表和测验焦虑问卷。这里主要介绍施皮尔伯格等人（1972，1978）编制的测验焦虑量表（Test Anxiety Inventory，简称 TAI)。

施皮尔伯格的 TAI 对焦虑的定义还是根据其状态——特质理论，把测验焦虑看成特质，看成个体的焦虑倾向性。他们把测验焦虑也分成两个部分、即 W 因素和 E 因素，测验焦虑特质高的人更倾向于把测验情境看成是对自我的威胁，因而在测验过程中常表现出紧张、忧虑、神经过敏及情绪冲动，从而分散注意力，干扰学生对智力认知任务的顺利完成。这里的 W 因素是指对失败结果的认知，而 E 因素则是由评价的紧张所引起的自主性神经系统的反应。

TAI 有 20 道题，要求被测验者报告他们在测验情境中通常的感受，按 4 种程度反应。例如：在测验中，我非常紧张

①从不；②有时；③经常；④总是。要求答题者根据自己的情况选择一个最适合自己的反应。测验可以个别或团体施测，没有时间限制，中学生和大学生都可在 8～10 分钟内填完表格。

该测验由叶仁敏（1990）[①] 作了修订，在上海市抽取了 535 人做样本，分别按性别建立了大学生、大学新生、业余职工大学生和高中生的常模，并分别建立了 TAI 总分、忧虑性（即 W 因素）、情绪性（即 E 因素）的常模。但缺乏信效度指标。

（四）其它临床焦虑量表

关于焦虑的临床量表很多，这里主要就《贝克焦虑量表》（Beck Anxiety Scale，简称 BAI）和《汉密顿焦虑量表》（Hamilton Anxiety Scale，简称 HAMA）作重点介绍。

1.《贝克焦虑量表》

《贝克焦虑量表》由美国 A. T. 贝克等人于 1985 年编制，适合于具有焦虑症状的成年人，主要是测量受测者主观感受到的焦虑程度。有研究表明，该量表亦适合于我国[②]。该量表有 21 个题目，采用 4 级计分方法，1 表示无焦虑症状烦恼，2 表示轻度（无多大烦恼），3 表示中度（尚能忍受），4 表示重度（只能勉强忍受）。其项目举例如下：

（1）腿部颤抖。

（2）头晕。

（3）手发抖。

计分方法较简单，只要把 21 题的总分相加，按 Y = INT(1.19X) 取整，转换成标准分即可，这里 X 表示总分粗分。

---

① 叶仁敏：《状态——特征焦虑量表（Y 版）指导手册》，1990 年出版。

② 《心理卫生评定量表》，《中国心理卫生》杂志社，1993 年（增刊）出版，第 191～225 页

其效度指标主要有两种：一是取60名焦虑症患者和80名健康人作BAI测查，对测验总分进行T检验，发现焦虑症患者得分显著高于健康人；二是对60名焦虑症患者用BAI和自我评定焦虑量表（Zung，1971年编制）进行检查，二者的相关为0.828。

2.《汉密顿焦虑量表》[①]。

《汉密顿焦虑量表》由汉密顿1959年编制，主要用于评定神经症和其他病人的焦虑严重程度。

HAMA与其它焦虑量表不同，它是由受过训练的评定员按照14个症状方面进行的5级评定（0－4，数值大表示严重），除第14项（即会谈时的行为表现）要结合观察外，其余项目都是根据病人的口头叙述进行评分，而且特别强调受测者的主观体验，其内容包括焦虑心境、紧张、害怕、失眠、躯体性焦虑、心血管系统等症状。每次评定，大约需10～15分钟。

根据全国精神科量表协作组的资料，总分超过29分，可能为严重焦虑；超过21分，肯定有明显焦虑；超过14分，肯定有焦虑；超过7分，可能有焦虑；7分以下便没有症状。一般来说，经过10次以上的训练，评定者有极好的一致性。上海市精神卫生中心曾对19例焦虑症患者作了联合检查，两个评定员之间的一致性很高，总分信度为0.93，单项症状信度为0.83～1.00；其实证效度也比较理想。

① 《心理卫生评定量表》，《中国心理卫生》杂志社，1993年（增刊）出版，第191～225页。

第二节　兴趣测验

一、兴趣测验概述

兴趣是个性的一部分，是人们从事各种活动的一种动力。一般将其定义成“积极探究某种事物的认识倾向”①，不同人的兴趣有不同的特点，这些差异表现在三个方面：一是兴趣的指向性差异，有的人对音乐感兴趣，有的人对体育感兴趣，有的人对哲学感兴趣；二是兴趣的广度差异，所谓广度是指的数量范围，有的人兴趣广泛，琴棋书画样样喜欢，有的人兴趣狭窄，除了自己的专业外，对其它内容一概不感兴趣；三是兴趣的稳定性差异，有的兴趣持续时间很短，有的兴趣是一辈子不变。一般而言，要进行测量的兴趣都不是短暂的，因为稳定性太差，测量的信效度难以保证。

兴趣测验通常要考虑两个基本问题：一是兴趣的客观表现，通常兴趣不是凭空存在的，它往往与一些活动分不开，如果一个人对体育感兴趣，他就会经常观看电视中的体育新闻，了解体育明星的经历和状况，学习体育比赛的知识，看体育杂志等；二是兴趣的主观表现，兴趣是一种主观愿望，有时仅仅通过活动了解是不够的，比如有的学生本不喜欢数学，但考虑

① 林信鼎等主编：《心理学辞典》，南昌，江西科学技术出版社出版，第177页

到数学成绩不好就考不上重点中学，为此他也可能刻苦学习，到处订数学辅导资料，找老师问数学问题等。只有主观上喜欢，并在客观上有所表现者，才能准确地判断其兴趣所在。

目前，心理测验学家对兴趣的研究很多，但主要集中在比较稳定的职业兴趣方面。职业兴趣测验的历史可以追溯到1927年，当时斯特朗（E.K.Strong）编制了《斯特朗职业兴趣调查表》（简称SVIB），此后，库德（G.F.Kuder）编制了《库德爱好记录表》。这两个量表都是严格按心理测量的要求构建的。与这两者不同的是霍兰德在50年代末编制的《职业爱好问卷》（简称VPI），他把职业兴趣分成6个领域，与职业兴趣相应，把职业也分成6个职业领域，可以根据被试的反应在职业分类表中确定职业兴趣。

除上述3种职业兴趣测验外，职业兴趣测验后期亦有一定的发展，但基本上没有什么实质性的突破，只不过是在做些完善工作而已。比如增加一些职业量表，增加问卷的有效性指标，寻求提高测验效度的办法等等，其中比较有影响的主要有白纳德（Brainard）《职业爱好问卷》、美国大学入学考试中心（简称ACT）《兴趣问卷》[①]、鲁尼波格（Lunneborg）（1968）编制的《职业兴趣问卷》（简称VII），限于篇幅，这里不作介绍。

---

① 龙立荣：《介绍国外四个著名的职业兴趣测验》，北京，《社会心理研究》，1991年第3期，第45－50页

二、常见的职业兴趣测验

(一)《斯特朗职业兴趣问卷》①

《斯特朗职业兴趣问卷》是世界上最早的兴趣问卷，它是根据经验编制的测验。其基本做法是这样的：取两组被试，一组代表专门从事某种工作而且喜欢该职业的所谓标准职业人员，而另一组则代表一般人，让两组受测者对测验项目进行诸如喜欢、无所谓和不喜欢的选择反应，由于这些人有差异，故回答不尽相同。斯特朗把这些能反映二者差异的项目合在一起，便构成某个标准职业的兴趣测验的项目集，不同的职业有不同的项目集组合（各职业有些项目相同），把这些不同的项目合在一起，就构成了该兴趣问卷的总项目。为了确定某个人的职业兴趣，将某人对所有项目的反应分别按各种职业标准量表计分，视其得分的高低，最终确定其职业兴趣。

由于《库德爱好记录表》的产生和发展，它产生的影响也越来越大，坎贝尔（D.Campbell）于1968年把库德量表中的同质性量表（比具体职业大的职业领域量表）引入了《斯特朗职业兴趣问卷》，另外在1972年，坎贝尔又把霍兰德的6大职业领域也引入了《斯特朗职业兴趣问卷》。这样，该量表结果便可以在三个层次上解释了：第一个层次为霍兰德的一般职业主题（简称GOT）；第二层为相互异质的同质性量表（简称BIS）；第三层为职业量表。下面就《斯特朗——坎贝尔兴趣问

① 龙立荣：《介绍国外四个著名的职业兴趣测验》，北京，《社会心理研究》，1991年第3期，第45～50页

卷》(简称 SCII) 1985 年版作一介绍。

SCII (1985) 有 325 个项目，这些项目的内容涉及职业、学校课程、活动、闲暇活动、人的类型等，要求对每个项目作喜欢、无所谓或不喜欢的回答，例如：

职业：	L (喜欢)	I (无所谓)	D (不喜欢)
电工	L	I	D
学校课程：			
天文学	L	I	D
闲暇活动：			
滚木球戏	L	I	D
活动：			
观看体育的电视	L	I	D

SCII (1985) 共有 264 个量表，包括 6 个 GOT 量表，23 个 BIS 量表，207 个职业量表 (代表 106 种职业)，2 个特殊量表 (学业满意度和内外倾量表)，26 个管理指标 (一是遗漏指标，二是奇特反应指标，三是对上述 7 个方面以及所有这些方面回答的百分数，共 24 个指标)。这里主要介绍职业量表、基本兴趣量表和一般职业主题的情况。

SCII 中的职业量表基本上是追循前面介绍的经验量表编制思路，在被测者答题后，分别按不同的职业量表计分标准计分，然后转化成常模分数，按常模分数的高低，确定受测者喜欢的职业和不喜欢的职业。BIS 的编制方法是把所有测验项目求两两相关，然后将高相关的项目合在一起。GOT 与上述二者不同，它是理论建构的，即先给各种类型下定义，然后再根据定义来确定每个类型的项目，每个同质的 GOT 量表共有 20 个项目。每个层次的嵌套是通过相关的办法确定的，即被包括的低一级量表都与 GOT 之间有高相关，其具体表现示例如下表：

表 15.2　SCII 的三种水平量表

GOT	BIS	职业量表
	农业	农民
	自然	
现实型	冒险	运动教练
	军事	
	机械	军官
	科学	地理学家、化学家
	数学	计算机程序专家
研究型	医学	牙科医生
	医疗服务	
	音乐/戏剧	乐师
艺术型	艺术	艺术教师、建筑师
	写作	律师
	教学	外语教师
	社会服务	学校领导
社会型	体育	
	持家艺术	家政经济学教师
	宗教	
	公开演讲	
	法律/政治	
企业型	经商	
	售货	
	商业管理	商店经理、投资经理
传统型	办公室事务	会计、银行工作人员 数学教师、秘书

根据被试测验的结果，将其放在所有职业量表、基本兴趣量表和一般职业主题上计分，即可得出该受测者的职业兴趣的总体状况。一般来说，职业量表中，如果标准分在 45 分以上，该受测者被认为与从事这一职业的人很类似，如果标准分低于 25 分，则认为与从事这一职业的人很不相似，而 26 分～44 分则被认为没有提供多少信息，最后给出很相似和很不相似的职业及分数。基本兴趣量表（23 个）和一般职业主题（6 个）的结果呈现如表 15.3 和 15.4 所示：

表 15.3　基本兴趣量表结果呈现示意

量表	标准分	说明
农业	39	相当低
自然	40	相当低
冒险	48	中等
教学	60	相当高
宗教	54	中等
办公室事务	40	相当低

表 15.4　一般职业主题量表结果呈现示意

量表	标准分	说明
现实型	38	稍低
研究型	46	中等
艺术型	57	中等
社会型	55	中等
企业型	51	中等
传统型	41	稍低

(二)《库德职业兴趣调查表》(简称 KOIS)[①]

库德于 1934 年编制了《库德爱好记录表》，其基本思想是：把所有职业分成 10 个兴趣领域，然后确定与之相应的 10 个同质性量表，受测者的结果按这 10 个量表记分，通过得分高低确定感兴趣或不感兴趣的职业领域。由于这种办法所测得的结果比较笼统，为此，库德从 SVIB 中吸取了职业量表的思想，在 1966 年编制了《库德职业兴趣调查表》(简称 KOIS)，1985 年，他再次修订了 KOIS. 这里就 1985 年修订版作些介绍。

KOIS (1985) 由 100 组 3 个项目构成的强迫选择项目组构成，这种形式可以避免反应定势，其形式举例如下：

项目	反应	
修理汽车马达	M (最喜欢)	L (最不喜欢)
计算平均成功率	M	L
挨家挨户卖杂志	M	L
在合唱队中唱歌	M	L
在医院做义务工	M	L
到森林中野营	M	L

在职业量表或大学专业量表的记分上，库德的记分办法与斯特朗不同，他主要是不取对照组，直接把个人的成绩与标准职业组或大学专业组的测验成绩进行比较，这里的大学专业量表是斯特朗量表所没有的。如果受测者与哪个标准职业组或大学专业组的分数接近，就说明其对该职业或专业感兴趣，确定感兴趣职业或专业的标准是最高相似系数之下相差 0.06 以内的职

---

① R.L.桑代克：《心理与教育的测量和评价 (下册)》，北京，人民教育出版社出版，第 120～132 页。

业或专业，一般呈现10个职业或专业。下表是某一个攻读心理测量学的年轻女性的结果：

表15.5 KOIS职业或大学专业兴趣测试结果[①]

职业量表女性常模		大学专业女性常模	
1. 心理学家	0.66	1. 心理学	0.61
2. 诊疗心理学家	0.63	2. 生物学	0.60
3. 计算机程序编制员	0.61	3. 英语	0.59
…………		4. 外语	0.58
4. 精神病社会服务工作者	0.59	5. 历史	0.55
5. 社会调查工作者	0.57	…………	
6. 书店经理	0.57	6. 保健	0.52
7. 保健治疗家	0.54	7. 数学	0.52
8. 医疗服务工作者	0.54	8. 普通社会科学	0.51
9. 中学理科教员	0.53	9. 基础教育	0.50
10. 学校社会服务工作者	0.53	10. 政治学	0.50

注：虚线以上为最感兴趣者。

除职业和大学专业量表外，KOIS还有职业兴趣评估和个人匹配部分。职业兴趣评估主要是过去的10个同质性量表，是对传统内容的一个修订本，按百分等级呈现结果，男女常模分开，分高、中、低三级职业兴趣领域，其标准为：高者为百分等级在75以上，低者为百分等级在25以下，中者在高低之间。

对于KOIS而言，在SCII中，由于其职业量表按经验法

① R.L.桑代克：《心理与教育的测量和评价（下册）》，北京，人民教育出版社出版，第120～132页。

建构，它对标准职业组的特点及其与其它职业组的区别很清楚，因此区分能力较强。而 KOIS 只研究各种职业的共同点，而且从事各种职业的人有许多共同之处，故一些相差很大的职业所得分数却比较接近。另外在同质性测验中，同质性职业内部还具有许多异质性，比如建筑师通常有绘画、制图、解决机械和数学问题的爱好，同时也有很多不同的爱好，如仔细考察建筑师的工作就会发现，这个人可能喜欢设计、建筑管理，而另一个却喜欢从事教学或建筑摄影。为了解决这个问题，库德提出了所谓个人匹配，即把一个人单独地与某职业中的不同个人样榜进行匹配，使对个人的兴趣进一步深入、具体。

（三）自我指导问卷

继斯特朗和库德之后，在兴趣问卷编制领域比较有建树的当属霍兰德。他从 50 年代开始进行这方面的研究。

1970 年，霍兰德编制了第一个《自我指导问卷》（Self—Directed Search，简称 SDS），1985 年又对其做了修订，这里简单介绍其内容。霍兰德的 SDS 主要由两部分构成，一是职业类型测验，另一是职业搜寻表。其基本思想是先测定自己的兴趣特性（也叫人格特点），然后根据自己的人格特点查找适合自己的职业。很显然，职业人格类型或特点与职业之间有一种内在的联系。

霍兰德把人格分成 6 种类型：现实型、研究型、艺术型、社会型、企业型、传统型。每个人的人格都是这 6 个维度按不同的程度组合而成的。与此相应，职业所需要的特性与这 6 个维度也密切相关。为了标定个人的兴趣特性或人格特性，霍兰德采用 3 个维度来标定。这 3 个维度的排列方式称为《职业三字母码》，如 RIA、ASE 等等。这样，经过第一部分测验所确定的三个字母码就可以和职业搜寻表中的三个字母码匹配了。

下面我们简单介绍一下这个测验的基本内容和实施过程。首先，要求根据个人的经历或感觉，确定自己感兴趣的职业，以便与后面测验的结果进行比较。第二步即是进行测量。这个测验有 4 个方面内容：活动、能力、职业和能力自我评价。每个方面的内容都按 6 种类型以 R－I－A－S－E－C 的顺序排列。而且每个方面的各种类型的题目的数目是相等的（能力自我评价例外，它主要是进行 6 种类型活动能力水平等级评估）。这些项目不是随机排列的，它是按 6 种类型分别集合在一起。下面是其测验部分的几个样例：

活动部分样例：

R	L	D	I	L	D
装修电器	L	D	研究科研课题	L	D
修理汽车	L	D	在实验室工作	L	D
上工艺课	L	D	上物理课	L	D

能力量表样例：

A	Y	N	S	Y	N
会一种乐器	Y	N	擅长向别人解释事情	Y	N
能够独唱	Y	N	能够做个好主人	Y	N
能制造陶器	Y	N	擅长判断人的性格	Y	N

职业量表样例：

E	Y	N	C	Y	N
旅馆经理	Y	N	记帐员	Y	N
推销员	Y	N	高校教师	Y	N
广告总经理	Y	N	税务专家	Y	N

注：L 代表喜欢、D 代表不喜欢、Y 代表有能力或喜欢、

N代表没能力或不喜欢

第三步即确定职业码。具体方法是这样的：把所有肯定的问答按6种类型记总分，取最大的3个维度按由大到小的顺序排列即可。第四步即根据这个职业三字母码在职业搜寻表中找职业，并将所选取的职业按自己喜欢的顺序来排列，因为每一类往往不止一个职业，而对职业的喜欢又有所不同。如果这些职业都不理想，则可以将三字母码重新排列，然后再在职业表中查找，这样，将喜欢的职业按顺序排列。一般说这些职业会与前面填的理想的职业基本一致。

第三节　态度和品德测量

一、态度测量

（一）态度概述

态度是指个体对人或事所持有的一种较为持久而又一致的心理倾向，它包括认识、情感和行动倾向三种成分。这三种成分起作用是有先后的，通常是认识在先，它的作用是形成对人或事物的了解、认识、看法，并在此基础上形成一定的评价，紧接着是情感，最后是意志行动倾向。有些时候，认识、情感、行动倾向是同步、和谐一致的，有时从认识到行动倾向却有一定距离。尽管态度相对稳定，但也不是不可改变的，比如教育、广告在许多情况下就是要改变人们的态度。

态度的准确评价至少有以下几种功能：一是了解人们对各种不同事物的态度；二是评价宣传工具在改变人们的态度中的效果；三是评价教育工作的成效。由于态度随人和事物的不同而存在着很大差异，因此态度测量更多的是提供一种科学的测量思路和方法，而不在乎形成某种固定的测验，这里主要介绍态度测量的几种常见方法。

（二）态度测量方法

1.《等距量表》法

这种方法由瑟斯顿 1929 年创立，故又叫《瑟斯顿量表》。他的基本思路是：围绕某一态度主题，选取能代表该方面的态度语或项目若干，由专家对这些项目进行等级排列，并把专家排列的结果进行项目分析，保留有效的项目以及根据专家的反应确定项目的等级。要了解某个受调查者的某方面态度，只需看其对该量表的反应，最后运用对全部项目反应结果（等级）求中位数，以中位数表示该受调查者的态度状况。这里比较困难的工作有两项，一是项目的收集和编制，二是项目的好坏及等级的确定。

（1）项目的编辑：项目编辑首先是要找到足够的态度语，一般在预试时要有 100～200 句。常用的编题办法有这样几条：第一是查阅相关的文献；第二是请来自不同团体的成员写出他们对特定事物的看法；第三是请相关问题的研究专家编写题目。在选题过程中，特别应注意找够中间等级的态度语句，通常两种极端的态度语比较多而且容易编。其次要使态度语的表达合乎以下几点要求：第一是措词简单，语义易于了解；第二是每一态度须针对本研究主题表示一个确切的态度。比如下面编辑的反应妇女在经济界的地位的项目就比较好：

● 结婚后，如夫妇各有工作，生活更为快乐。

● 妇女不应该依靠男子。

● 妇女的合适工作是管理家务。

(2) 确定项目的好坏及计分标准：为了确定上述项目的优劣与计分标准，常见的做法是请专家对前面编辑的项目进行等距排列，由最不赞成到最赞成，通常等级数不能太少，一般在7～13之间。如用1表示最不赞成，13表示最赞成，2表示不赞成，程度仅次1；12则表示赞成，程度仅次于13，其余类推。由于评定专家不止一人，因此评判的结果可能不一致，如何根据专家们的评判来决定项目的好坏和等级呢？假如这里有一个按11个等级排列的项目，各专家判断的等级的累计百分数如下图所示：

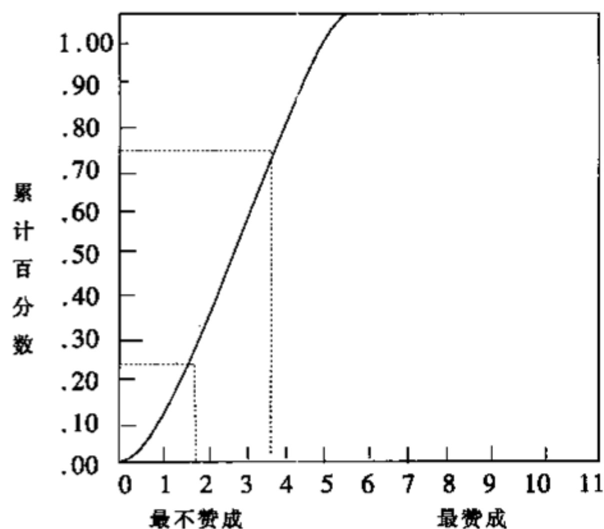


图 15.1　态度语评价的累计百分比分布

通过该图可以得到两个结果，一是该项目的量表值，二是该项目的鉴别力。项目的量表值是以项目累计分布的中位数（即50%累计百分比所对应的等级），而项目的鉴别力以Q值（四分差）表示，由累计百分图上的25%和75%的点所对应的等级Q_1和Q_3之差作为大小，即$Q = Q_3 - Q_1$。一般而言，Q值愈小，表示评判专家的态度愈一致，即态度语愈不含糊，质量好；Q值愈大，则说明该态度语愈不一致，质量差；Q值大于2的态度语应淘汰。

态度量表经由上述过程后，把合乎要求的态度语合在一起便构成了一个态度量表，这个量表的每个项目均有等级值。要知道某受测者的态度，只要求受测者作赞成与不赞成的回答。由于受测者的赞同反应不只一项，这就有一个如何估计受测者的态度等级的问题，通常的做法是把被测者表示同意的项目依分数高低排列，然后求出中位数，以居中项目的量表值作为该受测者的态度的估计值。

《瑟斯顿量表》的信度一般在0.8～0.9之间。《瑟斯顿量表》的不足主要有以下几点：第一是制定过程复杂，选项目、找专家评价都很困难；第二是用中位数代表态度等级不一定合适，因为中位数相同，但其余的反应未必一致；第三是项目的挑选和等级确定以专家的评判为依据，专家的意见能否代表一般人值得怀疑；第四是等距量表事实上是否真正等距，亦把握不准。尽管如此，《瑟斯顿量表》在主题比较清楚、调查范围不广的态度问题调查上效果还是比较好的。

2.《利克特量表》法

《利克特量表》法是由利克特（R.A.Likert）于1932年提出来的，是总加量表中最常用的一种。它的思路与《瑟斯顿量表》不同，在《瑟斯顿量表》中，认为量表中的每一个项目应尽可能反映不同的态度等级，因此主张请专家将项目分成等

级，《利克特量表》法则假定每一项目或态度语都具有同等的量值，项目之间，没有差别量值；另外受测者的反应也不相同，在《瑟斯顿量表》中，受测者只对态度量表中的项目答赞成与不赞成，而《利克特量表》中要求受测者对每一个项目的态度强弱按五级或六级反应；最后，在结果的估计上二者的办法也不相同，《瑟斯顿量表》用的是中数作代表，而《利克特量表》用的是受测者在所有项目中评定等级的总和。

《利克特量表》法的项目表述、等级评定和项目筛选上很有特色，这里着重介绍如下：

(1) 项目表述与等级评定：《利克特量表》在项目表述上有两种方式，即正面与负面陈述；而在等级评定上都是相同的等级数，只是在总计分上要考虑颠倒，保持标准同一，即负面陈述要把分数倒转。假如某态度测量为5级计分，非常同意得5分，同意得4分，无所谓3分，不同意得2分，非常不同意得1分，那么正面陈述的题目答非常同意得5分，而负面陈述的题则得1分，其余类推。

(2) 项目筛选：为了保证态度测量有效，保证每个题目的鉴别力是基础，那么如何来鉴别每个项目的区分度，通常的做法是将所有受测者的得分按总分由高到低排列，然后计算高分组与低分组在每一项目上的平均得分的差异，差异越大的项目鉴别力越好，反之则越差。

《利克特量表》的优点是制作过程简单，而且能广泛接受与态度主题有关的项目；另外可通过增加项目而提高效度，并且允许受测者充分表达态度的强烈程度。问题与不足是相同的态度分数者可能持有不同的态度模式，从总分只能看出一个人的赞成程度，而无法对态度差异作进一步的解释。

3.《哥特曼量表》法

《哥特曼量表》是由哥特曼（L.Guttman）于1950年提

出，这种量表的编制思路与前述二者不同，它试图确定一个单向性的量表，所谓单向性即项目之间的关系或排列方式是有序可循的，如果一个人赞同第二个项目，他同时也赞同第一个项目，如他赞成第三个项目，他也赞成第二、第一个项目。这种单向性是《瑟斯顿量表》所不具备的，尽管《瑟斯顿量表》中的项目有等级，但赞成高等级项目者未必赞成低等级项目。在《利克特量表》中，受测者的结果依项目总分而论，与单个项目的关系就更远了，正因为如此，瑟斯顿的中位数估计法与利克特的总分估计法对于相同分数等级的人都难于作出相同态度模式的测量结论。而《哥特曼量表》却有这种优势，相同分数的人，态度模式相同。

《哥特曼量表》的制定方法比较简单，现介绍如下：

(1) 挑选可用于测量对某事物态度的具体叙述句或称为项目，构成一个预备量表（假设有7个项目）。

(2) 将预备量表施测于一个有代表性的样组，赞成的项目以“0”表示，不赞成的项目以“×”表示（假设抽取了13人）。

(3) 将受测者按回答赞成的多少由高至低排列，将项目依赞成多少也由高至低排列，这样得到一个受测者对项目集的反应表，见表15.6。

(4) 去掉某些无法判断是赞成或反对的项目（这个假想数据未涉及）。

(5) 计算复制系数：复制系数的计算公式如下：

$$Crep = 1 - \frac{误答数}{总反应数}$$

它是单向性好坏的一个指标，如果复制系数高于0.90，则单向性得到基本保证。

那么何谓误答数、何谓总反应数呢?

总反应数为 13 个人，每人 7 次反应的总次数即 91；所谓误答数是指沿着答赞成与答不赞成的分切点所划的一条阶梯线(分切线上答不赞成或分切线下答赞成的即为误答数)，这些是不符合单向性标准的，从表中可知，不符合单向性模式的共有 4 个点，故 Crep = 1 - 4/91 = 0.96，属单向性比较好的哥特曼量表。将这些题目按新的顺序要求排列便得到了所需要的单向量表。

表 15.6　哥特曼量表反应分析表

被试	项目							分数
	7	5	1	2	4	6	3	
7	0	0	0	0	0	0	×	6
9	0	0	0	0	0	0	×	6
1	0	0	0	0	0	×	0	6
13	0	0	0	0	0	×	×	5
3	0	0	0	0	×	×	×	4
8	0	0	0	×	0	×	×	4
2	0	0	0	×	×	×	×	3
6	0	0	0	×	×	×	×	3
5	0	0	0	×	×	×	×	3
4	0	0	×	×	×	×	×	2
11	×	×	×	0	×	×	×	1
12	0	×	×	×	×	×	×	1

该量表的优点前面已经谈过，主要是由单向性带来的态度分数与态度结构的一致性，而缺点则是编制困难。

二、品德测量

（一）品德概述

品德是一个十分复杂的概念，从心理学的角度普遍的看法是，品德是一种个体现象，它是个人依据一定的道德行为准则，行动时所表现出来的稳固的倾向或特征，其结构包括道德认识、道德感、道德意志和道德行为方式；从教育学的角度看，比较流行的观点是把品德与思想品德等同，认为品德是一定社会思想、政治、道德的规范在个体身上的体现，认为品德的内容是思想品质、政治品质、道德品质的统一体。二者的差异是心理学对品德的内容定义得窄一些，而教育学要宽一些。另外，心理学更倾向于从过程考虑，而教育学从内容的结构与关系方面考虑多一些。这里则把品德的内容取教育学的观点，把品德的过程取心理学的观点，以利于后面对品德测量的全面介绍。

在个性中，品德是性格中能作善恶、好坏评价的主要内容，故它在人的个性中处于十分重要的地位；在教育目标中，德也是居于智育、美育、体育和劳动技术教育之首，它决定了人活动的方向以及价值，作用不可小视。科学准确地测量品德，不仅有利于检验教育的成效，而且有利于找到德育工作的成功经验和失败教训，以改进方式方法，最终达到接近教育目标的目的。

相当一部分品德测量在方法上比测验法宽，包含了观察法、实验法、访谈法乃至个案分析法等。不难设想，这类测量的信、效度不仅难于计量，而且也不会理想。鉴于上述这些原

因，这里不介绍这些内容，而主要择其中规范化程度高一些的情境性测验与问卷性测验作些介绍。

（二）情境测验法

情境测验就是设置一个活动环境或提出一个问题情境，通过学生对情境问题的反应，来了解品德特征。它分为直接情境测验和间接情境测验，直接情境主要是人为创造的真实活动情境，间接情境则是假想的问题情境。

1. 活动情境测验

这类情境通常是受测者需要亲自参加活动的情境，由于它比较具体，而且活动又不可能太复杂，因而只能了解品德的某一个方面，如诚实、公正、竞争与协作等，难于把品德的方方面面都反映出来。

哈特松（Hartshorne）和梅尔（May）是在品德研究中最早尝试情境测验的人。为了了解学龄儿童诸如诚实、自我控制和利他主义等品格，他们设计了一系列内容广泛的测验，其中应用最广的是诚实测验。其中一种方法是利用平常的考试情境，让学生完成一些诸如词汇、算术推理、完成句子一类的试题，考试完后，把试卷收齐带回，然后将试卷做一复份，下次上课时将未批改的试卷和标准答案发给学生，要求学生自己批改分数，再把批改后的卷子收回，将此卷与批改前的复份相对比，这样便可以发现儿童是否有自己修改答案提高分数的不诚实行为。

除此之外的常用情境是曲线迷、方迷和周迷三种情境，这三种情境的图形如下：

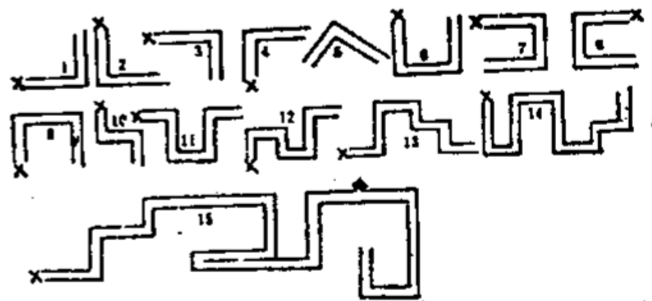


图 15.2　曲线迷测验

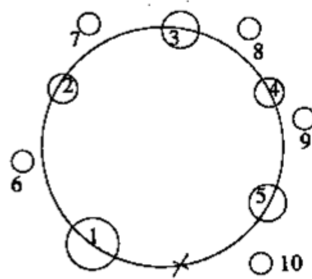


图 15.3　周迷测验

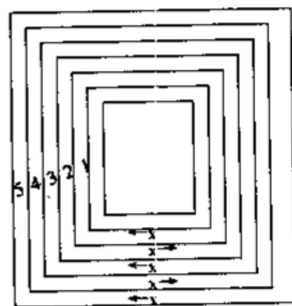


图 15.4　方迷测验

曲线迷测验要求受测者将铅笔尖放在迷津“×”处，同时要求受测者在闭上双眼的情况下按迷津的方向移动，并不可接触迷津的任何一边，完成一题得一分；方迷测验则是要求受测者在闭上眼睛后，从方迷的“×”处按箭头方向移动铅笔，也是不能接近边周的方框，最后回到“×”处，完成一个得1分；周迷测验也类似，只是要求在闭眼的情况下，从“×”处开始在大小不等的圆圈内打点，共打三遍，打中一个得1分。上述测验事先要通过控制测验确定诚实分数常模，然后将个人的操作成绩与常模分数进行比较，以确定诚实水平。

另外，前苏联心理学家雅可布松（C.г.яKobcoH）设计了一个测验公正的情境测验。其具体做法是这样的：把儿童分成三四人一组来玩玩具小汽车，让受测者一人隔着屏风负责分配各种不同的令人喜欢的小汽车，把儿童留给自己玩的小汽车数与分发给其他几个小朋友玩的小汽车数的比率作为公正性的指标，如果受测者平均分配，表明他公正，如果他的玩具数比别的儿童多，意味着他不公正。

2. 假设的问题情境

品德分为道德认识、道德感、道德意志和道德行为，活动情境的品德测验往往是一种综合测验，由于它涉及的面窄，因而对道德认识的发展是难于测量的。

美国心理学家柯尔伯格（L.Kohlberg）受皮亚杰（J.Piaget）的临床法及道德发展思想的影响，运用道德两难故事法这种假想的问题情境测验，间接地测量道德判断的方式及发展水平。他把道德发展分成3种水平6个阶段，即前世俗水平（preconventional level），它包含惩罚与服从的定向阶段和相对的实用主义阶段；世俗水平（conventional level），包括好孩子或好公民的定向阶段和遵从权威与维护社会秩序的定向阶段；后世俗水平（postconventional level），包括社会契约的定

向阶段和普遍道德原则的定向阶段。

他的道德两难故事中一个比较经典的故事是海因兹偷药。故事的内容大体如下：一个欧洲妇女患了癌症，生命垂危。医生告诉她的丈夫海因兹，本镇一位药剂师最近发明了一种镭化剂的药，可以救他的妻子。这位药剂师售一小剂药要价 2000 美元，高于药的成本 10 倍。海因兹竭尽全力只借到 1000 美元，他恳求药剂师便宜一些把药卖给他，或延期付款。但药剂师说："不行，我发明这种药就是为了赚钱。"海因兹在绝望中铤而走险，晚上去撬开药库偷了这种药。主试讲过故事后提了一系列问题问被试，如"海因兹该不该偷药？为什么？""法官该不该判他的罪？为什么？"这里主要是根据受测者对回答的理由及推理过程作发展水平的评价。

通常来说，在真实的情境与假设的情境中，对道德认识的测验是比较准确的，但在道德体验上是有差异的，实际情境要强烈得多，而且道德判断与道德行为之间往往并不一致，仅仅通过道德判断发展水平来预测行为是不完全的。

（三）问卷测量法

尽管用于测量人格、兴趣等的问卷很多，然而，专门用于测量品德的标准化的问卷却很少，造成这种状况可能有两方面的原因，一是品德问题太复杂，难于测量（特别是标准化地测量），二是过去对这个方面的研究尚不够深入。而前者的影响可能更大一些，因为品德的相当一部分内容涉及到价值判断，与人的生活密切相关，因而难于揭示，用自陈形式真实性就更难保证了。即使这般，只要能很好地设计，认真地控制，相信还是能反映一些客观信息的。

鉴于前述的种种原因，故这里只是简单介绍品德问卷测量的几种形式和注意事项。一般来说，问卷由两部分构成，一是

人口性资料，要求回答诸如姓名、性别、年龄、文化程度一类的问题；二是正式内容。问卷内容是两种呈现形式：一种叫封闭式，其答案都是规定好了的，受测者只需从中选择符合自己情况的一个答案；另一种是开放式的，不指出固定的内容，而让受测者围绕问题自由作答。为了保证结果的可比性、标准化和数量化，在问卷式测量中，封闭式反应形式用得较多，主要有以下几种：

（1）多项选择式：即要求受测者从固定的答案中选一个或几个答案。例如：

别人打你，你从不还手吗？　　A是　B否

我曾经说过假话　　A是　B不确定　C不是

（2）评定量表式：即要求受测者在每题后列出的几个等级中选出一个符合自己情况的等级。例如：

	很好	较好	中等	较差	很差
见了老师主动打招呼	1	2	3	4	5
诚实、不说谎、不骗人、不隐瞒	1	2	3	4	5
保护有益动物，不捉珍禽益兽	1	2	3	4	5

（3）排序或对偶比较式：即把一序列问题放在一起按符合自己情况的重要性排序或把这些问题两两对照，排出重要性。这种方式用得较少。

在品德测量类问卷的编制时，应注意以下问题：第一，应做好问卷的试用与修订。许多问卷在评价时，不考虑问卷本身的质量，不作同质性分析，不做信度和效度分析，难于使测量结果真实可靠；第二，要客观地看待品德问卷测验的结果，在某种意义上，品德测量比人格测量还难保证效度，因为它涉及的问题更敏感，如果把这种信息作参考那么它是有价值的，如果把它作为评价人的品德好坏的唯一标尺，那就会贻害无穷

练习与思考

1. 如何理解焦虑？常用的焦虑测量工具有哪些？

2. 职业兴趣测验的量表发展趋势是什么？如何评价职业兴趣测验在职业选择中的作用？

3. 态度和品德测量的常用方法有哪些？优缺点何在？

第十六章　测量的综合应用

本章提要：

● 测量在心理咨询中的应用
● 测量在人事测评中的应用
● 测量在教育评价中的应用

心理与教育测量的功能之一是评估人的特性，而人又是一个包含个性倾向性（如需要、兴趣、动机、价值观等）、能力、气质和性格等特征的有机整合体。因此尽管测量所应用的领域不同，但测量所涉及的内容却有很大的相似性，比如在心理咨询中就要测量人的各方面心理属性（这主要是为了帮助人更好地适应生活、社会），人事测评中也要测量人的各个方面的心理特点（这主要是为了选拔和安置合适的人），而教育评估中自然也少不了人的心理特性的评估（它主要是为了提高教育的效果）。因此介绍测量在这3个方面的应用时会有些交叉或重复，但由于需要不同，侧重点会有些差异。

第一节　测量在心理咨询中的应用

一、心理咨询概述

咨询① 一词来源于拉丁语 Consultation，其基本意思是商讨或协商、亦即通过商谈而解决问题。根据这种含义，在不同的领域有不同的协商，以解决不同的问题，如商业有商业咨询，法律有法律咨询，另外还有技术咨询、医学咨询和管理咨询等，心理学中也有以人的心理为内容而进行的心理咨询。心理咨询以人的心理方面的问题为内容，但不是任何人都需要心

① 邓明昱、郭念峰主编：《咨询心理学》，北京，中国科学技术出版社1992年9月出版，第3-4页。

理咨询，心理咨询的对象通常是那些存在心理冲突、心理不适、心理困惑或心理障碍的人。由于人们存在的心理问题的程度有轻有重，轻的如不知道如何与人交往，不知道如何对待失败，不了解自已适合干什么；重的如对某种动物有不同常人的恐惧，强迫重复某种不必要的行为、极端抑郁、神经衰弱乃至自杀念头强烈等。因此在“咨询”的过程、方法上存在一定的差异，正是这些差异，人们在对心理咨询的对象的认识上也存在一定的分歧，这些分歧的核心和焦点是咨询对象的心理困扰程度。

可以把心理困扰的状况形象地看成是一个连续体，如图16.1所示：

无困扰　　有轻度困扰　　有中度困扰　　有严重困扰

图16.1　心理困扰程度示意图

有人主张心理咨询的对象是所有存在心理困扰需要咨询的人，即无困扰之外的所有人。另有人主张心理咨询只针对有轻度困扰者，理由是有中度或严重心理困扰的人通常有神经症（如恐怖症、强迫症等）和精神病（癔病、性倒错、精神分裂症等）。他们的困扰的解除办法主要是心理治疗和药物治疗，而心理咨询只是辅助巩固疗效的办法，以治疗为核心，治疗者与被治疗者的关系不是咨询关系，而应是医患关系。因此建议把这类活动称为心理治疗更合适，它与一般意义上的咨询完全不同，故这部分对象不属于心理咨询的范围。第三种观点是折衷派，主张把轻度困扰者与部分中度困扰者划归为心理咨询的对象。他们认为第一种观点对心理咨询的定义过于宽泛，往往

容易夸大心理咨询的作用，而贻误对咨询对象施以及时的心理与药物治疗，会影响咨询的形象；而第二种观点又对心理咨询的作用估计不足，认为心理咨询对中度困扰者一筹莫展。事实上，许多中度心理困扰如恐怖症、中度强迫症等通过心理咨询亦可以收到很好的效果，而且这种方式相对于心理治疗更容易为咨询对象接受，实施咨询使咨询对象心理压力小，故主张在二者中打个折扣。

综上所述，可以对心理咨询作如下定义：心理咨询[①] 是求询者就其心理冲突、心理障碍或轻度心理疾病向有专业技术的咨询人员诉说、询问，咨询人员分析问题的原因和症结并寻求解决问题的办法，提高对生活的适应性和对周围环境的调节能力。

心理咨询的形式，从不同的角度有多种分法：按与接受咨询者的接近程度可分为直接面谈咨询和间接咨询（通过对咨询对象有关的人咨询而咨询）；按接受咨询者人数的多少可分为个别咨询和团体咨询（超过 1 人）；按进行咨询的手段分有电话咨询、信函咨询、电视咨询等。上面的各种咨询形式各有优缺点，在选择时要因地制宜。比如直接面谈咨询就有获得信息直接、影响直接和咨询效果好等优点。如果受咨询者胆小，或地处偏远，或经费短缺，直接面谈法的作用就难以发挥，信函、电话等咨询方式更有优势。

心理咨询通常有 3 个核心过程：一是分析诊断过程；二是帮助指导过程；三是效果评估过程。

分析诊断，就是要了解接受咨询的对象存在些什么困扰，有哪些具体的表现，导致这些现象的可能原因是什么，并在深入了解的基础上作出判断，确定问题属于哪种类型，程度严重

① 张小乔：《心理咨询治疗与测验》，北京，中国人民大学出版社 1993 年 4 月出版。

与否，原因为何，初步的应对措施是什么。

帮助指导过程即根据分析诊断的结果以及提出的对策，具体落实对策的过程。比如我们发现某学生的考试焦虑属情绪性的，拟通过放松训练予以缓解和消除，帮助指导过程便需要系统地给咨询对象讲清帮助工作的思路、办法、时间及注意事项，然后按这套措施一步一步地实施，在本例中便要讲清什么是放松训练，它与考试焦虑缓解的关系，放松训练如何进行，要多长时间，要达到什么效果等。在大多数情况下，帮助辅导过程比较顺利，也有些时候，可能由于诊断不准确、对策缺乏针对性，需要在实施中予以调整。何以知道效果良好或不佳呢？咨询效果评估是重要的一环。

咨询效果评估的办法有多种，测验亦是其中用得较多的一种。之所以说上述三个过程是心理咨询的核心过程，是因为不同的咨询形式可能有些特殊的地方，比如个别会谈等。

除了上述3个过程外，在开始还有一个初步的开端，以建立良好的信任关系，促成咨询活动的深入。

二、心理测量在心理咨询中的应用

如果把比内为鉴别智力低下儿童而编制的《比内智力测验》看成广义的咨询，那么心理咨询中应用心理测量的历史与心理测量的历史一样长。除智力测量外，兴趣、能力测验、焦虑测验，在心理障碍咨询中都得到了广泛应用。心理测量在心理咨询中的作用主要是诊断与效果评估，尤其以诊断用得最多。

(一) 在自我认识、人生规划咨询中的应用

从某种意义上说，人的一生都是在自我发现、自我创造。完全客观地认识自己并不容易，而不能较好地认识、评价自己，塑造自己就会很困难。

自我认识的内容很多，从自我认识咨询方面看，人们感兴趣的问题通常有两大类：一是自己是什么样的人，有什么长处，有哪些不足？二是这些长处适合从事什么职业，这些不足如何克服和弥补？由于这个问题与人事测评交叉太多，故放在下节介绍，这里只简单讨论第一个问题。

从评价自己的心理特性方面讲，人们比较感兴趣的主要是性格和智力以及价值观、气质类型等。一个人能较好地认识自己，比较客观地对待自己的短长，扬长避短，对自己有好处，对社会也有价值。假如某人发现自己在性格上乐群、外向、情绪稳定，但权宜敷衍、缺乏恒心和负责的精神，那么在以后的学习、生活和工作中便要保持优势，克服权宜敷衍的不足；如发现在能力上言语思维能力强，而动手能力、体育运动能力特别差，那就要注意克服不足，加强言语思维能力方面的优势，朝文学、法律、哲学、行政管理等方面发展。在性格测验方面，比较常用的测验有卡特尔人格问卷、YG性格问卷等。

《卡特尔16种人格因素测验》结果比较丰富，除了评价乐群性、聪慧性、稳定性、恃强性、兴奋性、有恒性、敢为性、敏感性、怀疑性、幻想性、世故性、忧虑性、实验性、独立性、自律性、紧张性这16个单维因素外，还有二元个性因素适应与焦虑型、内向与外向型、感情用事与安祥机警型 、怯懦与果断型的量表，以及对人的心理健康因素、有成就者的个性因素、有创造性的个性因素等方面的测量，16PF是咨询中广泛使用的人格测验。

《YG 性格测验》的使用也很多，它的特点是解释接近生活，容易理解，它把性格分成抑郁性（D）、稳定性（C）、自卑（I）、神经性（N）、客观性（O）、攻击性（Ag）、协调性（Co）、活动性（G）、适应性（R）、思维向性（T）、支配性（A）和社会向性（S）12 维度，然后又把这 12 个维度概括成情绪性、社会化、内外向性、活动性、冲动性、主导性等 6 个方面（具体如图 16.2 所示），该测验把人划分成 5 种典型类型和许多类型（或称混合型）。这 5 种典型类型是：适应外向性（D）、不适应冲动型（B）、均衡型（A）、适应内向型（C）、不适应内省型（E）。

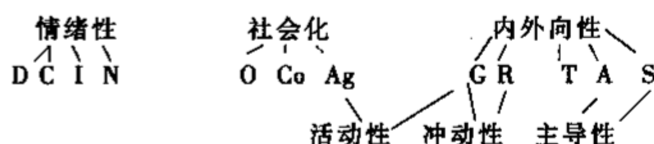


图 16.2　YG 性格维度与概括类别

《艾森克人格问卷》则适用面窄一些，主要用于临床，其测量结果主要有 3 个方面：内向还是外向，情绪稳定不稳定以及精神病症状。这些测验都有适用于不同年龄的常模（卡特尔人格因素问卷中适合 8～14 岁儿童的问卷简称 CPQ）。

在能力测验方面，比较成熟的有智力测验、能力倾向成套测验和单项特殊能力测验，这方面的内容在下一节要详谈，在此不多说。关于兴趣、价值观、气质等方面有一些测量工具，但标准化工作做得少，也不再介绍。

（二）在神经症、人格障碍等咨询中的应用

前面提到了对正常人或未出现明显心理障碍的人的性格的

测量，现在要讲在针对有中度或严重心理困扰者来咨询时使用的测量工具。

比较常见的心理障碍可分成 3 大类：一是神经症，包括恐怖症、强迫症、神经衰弱、焦虑和抑郁等；二是人格障碍，包括癔病、躁郁症、精神分裂症等；三是性心理障碍，包括恋物癖、窥淫癖、裸露癖和同性恋等。对于这一类前来咨询者，如有中度症状可通过心理咨询而治疗者，可以通过咨询而解除症状；如经过诊断超出了咨询的范围，可以推荐其到精神病医院接受治疗。

在这类咨询中，常用的诊断和评估工具有《明尼苏达多项人格问卷》（MMPI）、《艾森克人格问卷》、《症状量表（Scl—90)》以及部分焦虑测验量表。MMPI 是在临床诊断中比较权威的自陈测验，由于它是按照经验法编制，故在对咨询对象的症状与严重程度方面的评估比较准确，除了 4 个效度量表（疑问量表、说谎量表、诈病量表、校正量表）外，还有 10 个分量表，分别测量咨询对象在疑病症（Hs）、抑郁症（D）、癔病（Hy）、精神病态（Pd）、男性化 - 女性化（Mf）、妄想狂（Pa）、精神衰弱（Pt）、精神分裂（Sc）、轻躁狂（Ma）和内向（Si）方面的状况。此外还可根据这 10 个方面的两个高点组合，得到十几种组合评价，如果把全部的 566 题做完（前面的临床量表只需做 399 题），还可以对焦虑（A）、压抑（R）、外显性焦虑（MAS）、自我力量（Es）、依赖性（Dy）、支配性（Do）、社会责任感（Re）、偏见（Pr）、社会地位（St）、控制（Co）等性格内容进行测量。

《艾森克人格问卷》由 P、E、N 和 L 共 4 个量表组成，主要测量内外向（E）、情绪稳定性（N）、精神质（P）。L 量表是效度量表，主要测验受测者的不真实回答。在这个测验中，P 值在心理咨询中的作用比较大。

症状量表[①]（Symptom Check－List 90）是由德瑞格提斯（L.R.Derogatis）编制，起初这个量表仅适用于精神科或非精神科成年人门诊的病人，后来发现其对心理健康状况的研究亦十分有用。它包括90个项目，采取5级计分制方式，主要测量9个方面，即躯体化（主要反映身体不适感，如心血管、胃肠道、呼吸等系统的不适和头痛、腰痛、肌肉酸痛等）、强迫症状、人际关系敏感、抑郁、焦虑、敌对、偏执（主要指投射性思维、猜疑、妄想、夸大等）和精神病性，在咨询诊断与效果评估中有广泛的使用价值。

第二节　测量在人事测评中的应用

一、人事测评概述

人事测评是人事心理学中的一个核心问题。人事心理学致力于探讨人与事的最佳匹配，主要探讨人员选拔、训练、考核、分配和激励等有助于实现组织达到最大效率目标的规律。之所以说人事测评是核心，是因为组织目标的实现离不开人，而人的敬业精神、才能等心理品质只有和相应的事实现最佳的组配，才能真正有效地为实现最大效率创造必要条件，为管理的科学性打好基础，不然训练、考核、分配与奖励的效果会系

① 王征宇：《症状自评量表（Scl－90）》，《上海精神医学》，1984年2月第68～70页。

统地受影响。那么何谓人事测评呢？人事测评是指根据职业或工作的要求，通过各种测量手段，对人的素质适合事的程度进行评价的过程，其目标是实现人与事的最佳匹配。人事测评的基本假设是：事与事不同，人与人有差异，事不能不择人，人不可能适合干任何事，要充分发挥人的潜能，其条件之一就是要使人与事匹配。

人事测评的作用最早是在战争中显示出来的，在一战时期，美国人为了防止低能的和不合格的士兵入伍，于是便请心理学家编制了团体智力测验，为挑选聪明的士兵入伍和使聪明的人担任更重要的任务作出了贡献。受这种积极效果的激励，美国在二战时期还运用了按瑟斯顿思想编制的一般分类测验（General Classification Test，简称 GCT），按知觉速度、推理能力、语词理解和语词流畅、空间知觉、记忆和计算 7 种能力（即瑟斯顿所谓的智力），对军人进行分类，为战争的胜利作出了贡献。另外，美国 1942 年运用心理素质测验帮助挑选飞行员，结果使淘汰率由 65% 下降到 36%，大大节省了财力、人力的浪费。在 40 年代这种重视人员选拔的观点基本上是美国人事工作的常规事宜，可见其价值之重大。

人事测评工作有三步：一是工作分析，二是按工作分析的要求选人，三是选拔使用后的效果评估。

工作分析是人事测评的第一环，也是基础、关键的一环，它的好坏决定了选拔和效果评估。工作分析要解决两个方面的问题，即对工作本身作出规定和确定工作对工作人员的行为有什么要求。工作特性分析包括下列内容：①职务名称（便于工作登记、分类及确定组织内外的各种工作关系）；②工作活动和工作程序（说明所完成的任务、使用的原材料和机器设备、与他人工作的关系等）；③工作条件与物质环境；④社会环境（说明工作群体中的人数、完成工作所要求的人际交往、相互

作用的数量和程度等)；⑤雇佣条件（说明工作时数、工资结构、福利待遇、此工作在组织中的地位、晋升和调动等)。工作人员的行为要求是以工作特性分析为基础而制定的，其内容主要有：①身体方面的特性（包括体力、身体的灵活性、感官能力等)；②心理方面的特性（主要是一般的智力和创造力等)；③学习方面的特性（主要是表达能力、决策能力、知识与技能的学习能力等)；④激励方面的特性（主要指人格特性如志向、道德感、适应、自控、忍耐、孤独、依赖性等)；⑤社会方面的特性（主要指领导和协调方面的特性)。

人员选拔是继工作分析后因事择人的又一关键步骤，人员选拔的方法有多种，如面谈、工作申请表考察、心理测量等，对人的心理特点的测量是心理测量的优势，它可以发挥十分重要的作用，它的质量依赖于测验工具的优劣。

人员选拔的效果评估是对人事测评过程的一个检验，它的作用至少有以下几点：①判断预测或选拔工具的优劣；②帮助确定培训计划的目标；③给雇员提供具体的反馈。效果评估有主观测量和客观测量，主观测量可以用主观评定量表和定性评语评价，客观测量主要是生产数据和人事数据（如事故、离职、缺勤等)。

二、测量在人事测评中的应用

心理测量在人事测评中有一定的应用价值。如果把人员分成在岗与不在岗，那么对于在岗人员来讲，心理测量的应用有两个方面：一是在岗人员是否合格的诊断，二是对不合格者重新分配的工作安置及培训效果评估；对于要挑选的不在岗人员

而言，主要是选拔。如果把这两类人员合在一起，心理测量在人事测评时的应用主要有 3 个方面：一是人员的心理特点评估；二是人员培训后的心理特点评估；三是工作人员的绩效评估，下面分别介绍。

（一）在人的心理特点评估中的应用

不同的组织由于其结构、性质、规模等不同，自然其所要求的职位、工作也不一样，要使组织和谐、高效地运转，除了管理等因素外，选拔或安置合适的工作人员也十分重要。由于工作不同，其对人的心理特点的要求也就不一样。对人的心理特性的测量有两个大的方面：一是一般心理品质测量，主要指智力、个性等；二是专业知识和特殊能力测验。下面介绍几种常用的测验。

1. 智力测验

它在高级职员的选拔和安置中是经常用的指标，因为它是胜任这类工作所不可缺少的，一个智力低下得连自己的生活都安排不了，如何能领导一个工厂？指挥一项科研？常用的智力测验有《韦克斯勒智力测验》和《瑞文标准推理测验》。《韦克斯勒智力测验》的特点是把智力分成言语和操作两大块，又把言语部分分成常识、背数、词汇、算术、理解、类同 6 个分测验，把操作部分分成填图、图画排列、积木图案、拼图、数字符号 5 个分测验，个别施测，通常需要 45～60 分钟时间。《瑞文标准推理测验》则是非文字的标准化测验，有 5 个分测验，每个分测验 12 题，它可以个别施测，也可以团体施测，但无时间限制，一般成人 45 分钟左右可以完成。

2. 个性测量

不同工作对个性的要求是不同的，有些工作是单调重复的，要有忍耐力；有些工作需要与人打交道，需要外向的人；

有的工作在整个生产中十分关键，要求严格，压力大，要求人能忍受压力；有的工作有很大的风险，要求人有冒险性。评价人的个性，使用个性测验是一种较好的评价方法。在个性测量中，通常的分类是自陈测验，投射测验、情境测验和评定量表。自陈量表前面介绍较多，主要有兴趣、气质、性格等几类。在兴趣测验中，有《斯特朗——坎贝尔兴趣问卷》、《库德兴趣调查表》、《自我兴趣测验》；在人格（含气质、性格）测验中，有《卡特尔 16 种人格因素测验》、《艾森克人格问卷》、《YG 性格测验》等常用工具。投射测验主要有《罗夏克墨迹测验》和《主题统觉测验》，由于过程复杂、要求高，所用不多。情境测验、评定量表编制得还不多，这里不作介绍。

3. 专业知识技能测验

它在各类专业人员的选拔和安置中应用广泛。在许多情况下，仅仅测量智力、个性是不够的，因为智力、个性等代表人的一般素质，但如果需要的是有专业知识、技能的人，这就要求进行专业知识技能的评估。比如目前进行的公务员资格考试、会计师资格考试、计算机程序员考试、律师资格考试等均属于此系列。

4. 特殊能力测验

除了前面介绍的几种指标外，关于特殊能力的研究亦十分广泛，因为许多职业的专门化程度都比较高，它也许对一般能力的要求并不高，但特殊能力却必须达到基本要求，下面分别介绍：

（1）音乐能力测验：美国音乐心理学家西塞尔（E. Seashore）在依阿华大学进行了广泛研究，并编制了《西塞尔音乐才能测验》（The Seashore Measures of Musical Talents），这种测验后来几经修订，现在的形式是由 6 个分测验组成，包括音高、音强、节奏、节拍、音色和音高记忆。英国

心理学家维因（Wing）[①] 编制了《维因音乐智力标准测验》（The Wing Standardized Test of Musical Intelligence），测验有7个部分，前3个部分测量感觉辨别力；后面4个主要测受测者在比较两支曲子的美学优点上的欣赏能力，这7个部分是和音分析、音高变化、记忆、节拍重音、和声、强度、乐音的分节。

（2）美术能力测验：美术能力主要包括两个方面，一是艺术鉴赏能力，一是创作能力。一个没有绘画才能的人可能有较好的艺术评价能力，而一个画家则二者缺一不可。在艺术鉴赏方面，《梅尔美术鉴赏力测验》（Meier Art Judgement Test）是一个比较著名的测验，这个测验收集了许多名画，把名画与名画的改动版进行艺术感受评价，以鉴别人的艺术鉴赏力。在艺术创作能力方面，有《洪恩艺术性向量表》[②]（The Horn Art Aptitude Inventory），这个测验主要采用临摹作品的方法，它需要高度的创造力。测量内容包括素描画、随意画（要求受测者用指定的图形画出简单的抽象图案）、想象画（给受测者12张印有几条线的卡片，要求受测者用这些线条画成一幅草图），用优、中、差三级计分，以标准样图作依据决定好坏，最后作出总体评价。

（3）文书能力测验：有些文职人员经常要与数字、字母、文字等打交道，哪些人能干得又快又好，可以通过文书能力测验来判断。适用于这类能力检测的测验有《明尼苏达文书测验》、《一般文书测验》等。

（4）机械能力测验：与机械打交道的工种很多，但尽管如

---

① 黄元龄：《心理及教育测验的理论与方法》，台湾，大中国图书公司印行，1987年1月出版，第139页。

② 黄元龄：《心理及教育测验的理论与方法》，台湾，大中国图书公司印行，1987年1月出版，第316页

此，它们对人的要求却有很多的相似性，判断一个人机械能力的好坏，主要从机械知识、运动反应速度、协调性等几个方面测试，这方面的测验工具较多，主要集中在机械知识和手指灵巧、眼手协调上。比如《贝内特机械理解测验》(Bennett Mechanical Comprehension Test) 就是用来测机械知识的，它的内容主要选择日常生活中涉及机械原理的情境，要求人们利用相关的原理作出判断；而《珀杜插栓板测验》(Purdue Pegboard) 则主要是测人的手工灵活性和协调性，它有两部分，一是要求受测者分别用左右手将大头针插入小孔，二是把大头针、小环和橡皮圈装配到每一个小孔中（这时对手不作限定操作）；而《奥康纳手指灵活性和镊子灵活性测验》(O'Connor Finger and Tweezer Dextevity Test) 则主要测手指的灵巧性，它要求受测者尽快用手或镊子把针插入小孔中，这种能力对缝纫机操作员、牙科工作人员是不可少的。

(5) 多项能力倾向测验：这种测验在人的特殊能力的全面诊断评估方面很有好处，相对于前面介绍的一些特殊能力如文书、机械等有它的优势，其缺点是费时。目前介绍、修订较多的主要是一般能力倾向成套测验 (General Aptitude Test Battery，简称 GATB)，它包括普通推理能力、语言能力、数学能力、空间关系能力、形状知觉能力、文书能力、动作协调、手指灵巧、手工灵巧等分测验。

(6) 管理能力测量：由于管理工作十分复杂，要求也比较高，故对管理能力的评价也较困难。除了前面介绍的智力测验外，主要有两种测量办法：一是情境测验，二是评价中心方法。情境测验用得较多的是无领导群体讨论和《文件框测验》[1]，前者是让受测者在一定时期内就某些论题进行讨论，

---

[1] 韦恩·卡西欧：《人事心理学》，北京，中国人民大学出版社，1991 年 5 月出版，第 279～285 页。

这些人中没有领导，然后由评定者根据各种标准对受测者的表现进行评价；后者则是根据所要求的管理者的类型确定管理者的管理能力特点，然后给受测者类似于工作情境的任务，要受测者完成，最后由评判者按事先规定的标准对受测者的操作进行评分。这种测验的信、效度都比较高。评价中心方法① 吸收了问卷测验与情境测验的优点，对影响管理水平的7种评价要素（即行政管理技能、人际关系技能、智力、绩效的稳定性、以工作任务为中心的激励能力等）进行系统的评价。严格地说它是一种综合测量方法，其特色是它规定的7种能力要素，测量工具包括智力测验和文件框测验等。

（二）人员培训后的心理特点评估

在许多情况下，选拔的新员工上岗前或老员工的重新安置前都要进行专门性的培训，以使他们尽快提高工作技能、了解工作任务，干好工作，那么培训或选拔安置的有效性是一个值得考虑的问题。培训效果与培训目标是否一致，就可以用测量的办法，这里测量的可以是知识、技能水平的提高，也可以是工作态度、工作兴趣的改变，这些内容都可以用成就测验、兴趣或态度测量的办法进行评估。

（三）工作人员的效绩评估

对工作人员进行效绩评估既是生产管理的必要措施，也是人事管理的重要依据，这里主要介绍对领导者行为效果及员工心态评估的PM量表②。PM理论由日本大阪大学心理学家三

---

① 韦恩·卡西欧：《人事心理学》，北京，中国人民大学出版社，1991年5月出版，第279～285页。

② 徐联包、凌文輇主编：《组织管理心理学》，北京，北京科学技术出版社，1988年出版，第348～356页

隅二不二在九州大学任教时提出。他认为任何一个团体都具有两种机能：一是团体的目标达成机能，二是维持强化团体或组织体的机能，前一种机能简称为 P（performance）即工作绩效，后一种机能简称为 M（maintenance），即团体维系。领导者的作用就在于执行这两种职能，领导者的行为也就包括这两个因素，如果以 P 为横坐标，M 为纵坐标，并在 P 和 M 的中点各画一条平行线，领导者类型就可分为 4 种（如图 16.3 所示），其中 PM 型最好，pm 型最差，P 型和 M 型居中。如何评价领导者的行为类型以及工作效果？他根据广泛调查，从数百个有关问题中，通过项目分析和因素分析方法获得了 60 个题目，构成了 PM 问卷调查表，它由 P 因素量表、M 因素量表和情境因素量表构成，其中 P、M 量表各 10 个题目，8 个情境因素各 5 个题目，这 8 个情境因素是：对工作的意欲，对待遇的满意程度，对公司的满意程度，心理保健，集体工作精神，会议成效，信息沟通，绩效规范。这个评价可以由领导者自己评，也可由下级评。由 P、M 量表的得分区分领导类型，用情境量表的得分高低作为部下士气、态度和满意度的反映，它也是领导效果的表证。

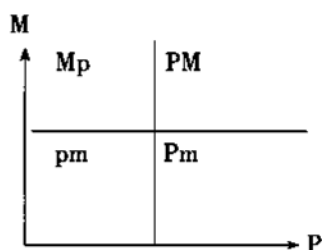


图 16.3　PM 领导的 4 种类型

第三节　测量在教育评价中的应用

一、教育评价概述

谈测量在教育评价中的应用之前，弄清教育评价本身的含义是十分必要的，而评价又是教育评价的基础，故我们先来分析评价。与评价关系最密切的有测量、测验和考试等。测量是"根据一定的法则用数字对事物加以确定"，测量的方法和内容十分广泛，仅就方法而言，就有以观察、实验、访谈等进行的测量和用测验法进行的测量。测验只是测量的一种方法，故它比测量的范围要窄。测验通常是指测量一个行为样本的系统程序，它的标准化程度比较高，而且受信效度指标的制约，这是相当一部分测量方法所不具备的属性。而考试从内容上讲，与测量的内容范围同样广泛，但从方法上讲与测验更相似，其与测验的区别主要在于考试多用于有目的的人员的甄别选拔，就目前的现实而言标准化程度比较低；而测验不限于选拔，还有鉴别个别差异、建构理论等作用。总之，测量的内容和方法是包容性均较大的一个概念，包含测验与考试。

那么什么是评价呢？通常认为评价是"对测量的结果进行价值判定"，测量是评价的前提，是评价的必要组成部分，没有测量，价值判断就无法进行。

在明确了评价之后，我们再来谈谈教育评价。评价是多方面的，可以是教育，也可以是心理疾病的治疗效果，还可以是

企业管理水平等。那么什么是教育评价呢？教育评价[1]是根据教育目标，在系统收集资料的基础上，对教育过程及其结果进行价值判断的过程。首先，教育评价要有一个参照标准，这个标准就是教育目标，教育工作的好坏主要应根据教育所能接近教育目标的程度；其次，教育评价要以事实为基础，没有一整套切实可行的科学程序，得不到符合客观实际的事实，价值判断就会出错误，而心理与教育测量的一整套理论和以这套理论为依据开发的成果对于获得客观的事实是很有帮助的；第三，教育评价的内容是对教育过程及结果的评价，是一种动态性评价。教育评价的内容有广义、狭义之分，广义的评价内容包括宏观的内容（如教育制度、教育规划、教育投资、教育环境质量等等）和微观的内容（如学校教育目标、课程设置、课堂教学质量等），而狭义的评价内容则主要以教师和学生为对象，是对学校教育活动和学生发展质量的评价，本书主要取狭义评价内容。

要确定一项工作的好坏，评价是基本的。那么教育评价有什么功能呢？其功能大致可以概括为 3 点：一是导向功能，由于教育评价的标准是教育目标，这个标准的确立为人们指明了努力方向，一旦经过评价发现偏离了教育目标，人们就会主动调整自己的行为，向符合目标的方向前进；二是管理功能，科学的教育评价，可使人们明确自己的现状和职责，一旦人们发现自己的不足，就会想办法予以改进，从客观效果上讲，达到了激发人的动机，调动人的积极性的效果；三是诊断和选拔功能，这是教育评价的基本功能，教育评价可以使我们了解人的德、能、勤、绩等多个方面的情况，是进一步施加影响或进行选拔的依据。

---

① 瞿天山：《教育评价学》，武汉，武汉工业大学出版社 1992 年出版，第 12 页

教育评价过程通常分为3个阶段：第一是确立明确的可操作的教育目标，作为对事实进行判断的标尺，这个标尺的明确与否、正确与否会直接影响评价结果，比如现在主要注重素质教育目标，如果还以应试教育的观点为出发点，那么评价的标准就会发生变化，这是在教育评价之前应认真考虑的；第二是根据教育目标选择或编制可以测量这些教育目标的工具或方法，选择或编制测量工具或方法应以测量理论为指导，力求使用最有效的手段来获取事实资料；第三是通过测量所收集的资料和数据，对照教育目标形成一个价值判断。心理与教育测量在教育评价中的应用主要体现在如何选择或编制科学的工具来测量事物，为进行价值判断作准备、打基础。

二、测量在教育评价中的应用

测量在这里主要是对人的测量，在教育过程中的人主要有3部分：学生、教师和管理者，下面分别就这3类对象的测量作些介绍。

（一）在测量学生的学习与发展状况中的应用

评价学生的学习与发展，在教育评价中居于主导地位，它至少有3个方面的作用：

（1）摸清学生的学习和发展状况，是因材施教的前提。任何一种成功的教育，如果不是建立在尊重学生已有的学习和发展状况的基础上是不可思议的。班主任做思想工作要了解学生，科任老师要教好课要了解学生，要培养学生的健全人格要了解学生。

(2) 弄清学生的学习和发展状况，是评价教育过程中不同阶段成效的依据，比如在单元、期中、期末学习后，为了检验教育工作的好坏，便少不了全面了解学生的学习与发展状况，它是前一段工作的结束，又是进一步进行教育工作的基础。

(3) 弄清学生的学习和发展状况，是评价一种新的教育思想、新的教育措施、新的教育技术等有效与否的重要指标。一项工作，如果只是因袭传统就会僵化，只有不断研究、改革才能创新，也才有生命力。为了不断提高教育工作的水平，进行各种教育研究是不可少的，而任何一项方法、措施的终级目标都是塑造、培养人，这种探索的评价离不开学生的学习和发展状况。

对学生的学习和发展状况的评价主要有这样几个方面：学生的品德、学习能力、创造力、学习成绩、职业兴趣、性格、气质、心理健康状况等。下面分别介绍：

1. 品德测量

前面一章的品德测量中，介绍了品德测量中标准化程度比较高的两种思路：一是情境性测验，如哈特松等的诚实测验以及柯尔伯格的道德发展水平测验；二是问卷测验。除了自评外，还可以通过他人如家长、教师、同学等进行评估，不论哪种方法，严格地说，对品德这种复杂的心理品质的测量都显得太简单、粗糙、片面。尽管如此，这些测量办法都能部分地用于评价品德。

2. 学习动机测验

一般而言学习动机是学习的直接动力，在学习动机中研究较多的是成就动机，通常把它分成追求成功的动机和避免失败的动机，另外对成败归因的方式亦很重要。

这里介绍两个测验：一是叶任敏等（1988）修订的由吉斯米（J. Gjesme）与尼加德（R. Nygard）1970 年编制的《成就

动机量表》(简称 AMS)。它主要用来测量人追求成功的动机和避免失败的动机，在编制者看来，成就动机强的人向往成功、有自信心，对成绩感到骄傲，喜欢富于冒险性、挑战性和难度高、充分发挥个人能力的工作，而且对失败并不感到特别得在意和羞愧。该量表共有 30 道题，前面 15 道题测量追求成功的动机，后面 15 题测量避免失败的动机，有适合中学生和大学生的常模，个别、团体施测均可，5~10 分钟即可完成。

另一份成就动机测验由周步诚等 (1991)[①] 修订，主要测量成就动机、考试焦虑、成败归因、要求水平 4 个方面。这个测验把成就动机看成因活动性质而异，它包括知识学习方面的成就动机、图画和美工音乐等方面的成就动机。成败归因有两类：一是外部归因，一是内部归因，内部归因者学习动机强，外部归因者学习动机弱。这里的要求水平是指在假设的情境中个人期望完成任务的水平，认为成功动机强的人的要求水平高，回避失败动机强的人的要求水平低。这个测验适合于小学四年级至高中三年级的学生，每个年级小学段、初中段、高中段都有常模，该测验的分半信度为 0.83~0.89，重测信度为 0.79~0.86，由以学习成绩作效标的评价看，效度也比较满意。

3. 学习适应性测验

学习适应性是一种学习适应能力，是指克服困难取得较好学习效果的一种倾向，学习适应性包括热情、有计划地学习、听课方法、读书和记笔记的方法、记忆和思考的方法、应试方法、学习环境等。周步诚等 (1991)[②] 修订了一个学习适应性测验，适用于小学一年级至高三的学生，为了使测验有针对性，不同的年级段有不同的内容，其年级段有小学一、二年级

① 周步诚：《学习动机测验指导手册》，1991 年出版。
② 周步诚：《学习适应性测验指导手册》，1991 年出版。

段，小学三、四年级段，小学五、六年级段和初中与高中段。其测验的主要内容有学习态度、学习方法、学习环境等，另外还专门设立了“回答一致性”这一回答有效性指标。该测验的分半信度为0.71～0.86，重测信度为0.75～0.88，用成绩优秀和成绩差的学生在分测验和总测验上的反应结果进行差异考验，发现有预期的差异，表明有一定建构效度。

4. 智力测验

智力是影响学生学习的重要因素，也是教育培养的目标，在教育评价中经常遇到。适合于评价学生智力的测验较多，这里主要推荐3种：一种是《韦克斯勒智力测验》，有儿童智力测验（适合于6～16岁儿童）和幼儿智力测验（适合于4～6.5岁儿童），前一种由北京师范学院林传鼎和北京师范大学张厚粲主修，后一种由湖南医科大学龚耀先主修。第二种是《中国比内测验》（适合于2～18岁的儿童和青少年），由北京大学吴天敏修订。第三种是《瑞文标准推理测验》（适合于5～70岁的人），其特点是它主要由图画构成，对文化知识的要求较低，还可以进行团体测验。这个测验有两个修订本，一个由北京师范大学张厚粲等修订，一个由华东师范大学李丹等修订。

5. 性向测验

它主要用于了解学生的潜在优势，即经过同等训练个人的相对优势。目前经过修订的测验有《一般能力倾向成套测验》（GATB）的两个修订本，一个是由上海市教育科学研究所高德建、顾天祯等修订的《中学生一般能力倾向成套测验》（简称SS－GATB）①，测量学生的9种能力，即一般智力（G）、言语能力（V）、数理能力（N）、空间判断（S）、图形知觉

① 高德建等：《中学生一般能力倾向成套测验指导手册》，1988年出版。

(P)、符号知觉（Q)、运动协调（K)、手指灵巧（F）和手工灵巧（M)。这些分量表可以分成三类：学习能力（包括一般智力、言语能力和数理能力)、知觉能力（包括空间判断、图形知觉和符号知觉）和操作能力（包括运动协调、手指灵巧和手工灵巧)，样本主要是上海市中学生，有初一至高二5个年级的常模，量表由百分位数计分。另一个是由华东师大心理系戴忠恒[①]修订。这个修订版与前面一个主要不同的是：①它们的样本范围不同。前者为上海市样本，后者为全国十几个省市、自治区、直辖市的17个中等以上城市。②它们的对象不同。前者为初一至高二5个年级，后者为初二至高三5个年级。③它们的修订蓝本不同。前者以美国版为主，后者主要以GATB日本1983年第四次修订版为主。除了上述成套性向测验外，还有单项测验，如音乐、美术、文书、机械能力等性向测验。这部分内容可参见上一节有关部分。

6. 创造力测验

创造力是现代教育的中心目标之一，在这方面比较知名的测验主要是《托伦斯创造思维测验》和《南加利福尼亚大学测验》，这里不多介绍，相关内容参见第十二章第三节。

7. 学习能力测验

林传鼎和张厚粲等根据澳大利亚教育学会制订的《学习能力测验》修订了适合于我国小学四、五年级和初中一、二年的《少年儿童学习能力测验》，该测验是一种团体测验，由3个分测验组成，即：①找同义词；②算术推理；③语言类比。该测验主要测试了北京、天津和西安三地的1080名小学四、五年级和初中一、二年级的学生，测验的分半信度在0.62～0.90之间，测验成绩与语文、数学的相关为0.33和0.48，说明该

① 戴忠恒：《一般能力倾向成套测验简介及其中国试用常模的修订》，《心理科学》，1994年1月出版，第16～20页。

测验是有效的，但尚需进一步扩大测验题目，更广泛地取样，取得更大范围的常模。

8. 学习成绩测验

学习成绩测验是平时用得最多、最普遍的一种工具，然而就我国目前的标准化程度来讲，水平也是比较低的，主观随意性比较大，对具体内容亦没有开发出相应的建立了一定信、效度的工具。我们应该在平时的工作中，不断总结，积累材料，使之逐步走向规范化、科学化的轨道。

9. 职业兴趣测验

随着我国人才市场的逐步建立，人事管理的逐步规范化，尊重人的心理特点，考虑人的个人兴趣是大势所趋，在中等教育中，学生有两次分流，如何了解学生的兴趣，使他们学习他们喜欢的专业或职业，是充分调动人的积极性和挖掘人的潜力的关键。国内这方面的工作正在逐步展开，在没有现成的职业兴趣测验工具之前，引进和修订国外已有的测验是一个省时省力的办法。上海市进行职业辅导时，就借鉴了在世界范围具有广泛适用性的 SDS（参见上一章第二节有关内容），效果尚可。

10. 个性测验

这方面的测验修订和编制是比较多的，从适育评价的测验看，占主导地位的还是自陈问卷，投射测验和情境测验编制和修订得很少。从内容上讲个性测验中与教育评价有关的可分成两个方面：一是正常者的个性，二是不健康或病态者的个性，当然这里也不排除有些个性问卷中包括部分不健康或病态个性，但以正常个性为主。从正常个性测验方面看，目前主要有《卡特尔 16 种人格因素问卷》、《儿童人格问卷》（CPQ）、《YG 性格问卷》等；从诊断心理不健康或病态个性方面看，主要有《艾森克人格问卷》（儿童）、《症状自评量表》、《心理健康诊断测验》（MHT）等。除《心理健康诊断测验》外，其余的测验

前面都介绍过，这里简单介绍一下这个测验。这个测验是根据日本铃木清等人编制的“不安倾向诊断测验”修订而成的，适合于我国中小学生心理健康状况的诊断。该测验可以团体施测，适合于小学四年级至高中三年级的学生，测验由8个分量表构成，即学习焦虑、对人焦虑、孤独倾向、自责倾向、过敏倾向、身体症状、恐怖倾向和冲动倾向，该测验没有效度量表(即说谎量表)，测验的解释分总体解释和分量表解释。该测验的分半信度五年级到初中三年级以及高中二年级在0.84～0.88之间，全量表的分半信度为0.91，重测信度系数在两个月之后进行，上述5个年级的重测信度系数在0.667～0.863之间，信度较高；从效度方面看，该测验与《明尼苏达多项人格问卷》相关量表的相关为0.59，对精神科医生诊断为有神经症或精神病者施测的结果表明，其有一定的一致性，各分量表之间的相关大多数不到0.40，而且各内容分量表与总分的相关在0.536～0.70之间，表明有一定的结构效度，总之该测验的效度也是比较理想的。

（二）测量在教师与管理者评价中的应用

对教师的评价主要有这样几个方面，一是教师的资格评定，即教师的专业知识水平是否达到基本要求，专业知识包括文化知识和教育心理学方面知识；二是教师的教学艺术水平的评定；即教师的教学能力；三是教师的管理水平的评定，即教师在学生班级管理方面的能力；四是教师的个性评定。其中，教师的资格和教学艺术水平是其中的核心内容。在教师评定方面，虽说有一些办法和措施，但标准化水平还不高，还有待进一步研究。

对于教育管理者的评价，也不是十分系统、成熟，可参照第二节的有关内容予以评价。

练习与思考

1. 心理与教育测量在心理咨询中有哪些主要的应用领域？有哪些测验可资利用？

2. 心理与教育测量在人事测评中有哪些主要应用领域？有哪些测验可资利用？

3. 心理与教育测量在教育评价中有哪些主要应用领域？有哪些测验可资利用？

4*. 我国心理与教育测量在三种应用领域还有哪些要完善或填补空白的地方？

第十七章　心理与教育测量理论的新发展

本章提要：

- 经典测量理论的缺陷
- 潜在特质理论
- 项目特征函数与特征曲线
- 项目参数和被试能力估计
- 项目反应理论的优良性质
- 项目反应理论的应用
- 概化理论的基本思想及测验情境关系说
- 测验设计
- G 研究与 D 研究

第一节　项目反应理论简介

一、经典测量理论的局限

经过前十六章的学习，我们对建立在真分数理论基础上的经典测量理论（Classical Test Theory，简记为 CTT）已经有了比较详细的了解，甚至已在测验的实践中对它有了比较深刻的认识。历史上，经典测量理论无论是在理论的基础研究方面还是在实践的指导方面，均为心理与教育测验的发展作出了巨大贡献。当今，经典测量理论在测量研究中仍然占据着非常重要的地位，继续指导着多种测验的编制和应用，我们不能轻视对经典测量理论的学习和研究。

但是，经典测量理论的理论框架是有先天缺陷的，在测验实践飞速发展的今天，已日益显示出它的局限性。

第一是经典测量论理的信度估计精度不高。根据真分数理论假设，测验原始分数 X 线性分解为测验真分数 T 和误差分数 E 两部分，并且进一步假设真分数是测验原始分数的期望，误差分数与真分数相互独立，从而导出测验信度为真分数方差与原始分数方差之比。且不说这一连串假设的可靠性，就说这结果，如此定义的测验信度并无助于信度的估计，因为在定义中除原始分数方差可得之外，真分数方差与误差分数方差都是无从求取的。为实际估计测验信度，经典理论又提出了平行测验概念或者条件稍弱一点的 τ 等价测验概念，从而推演出若干

信度估计公式。但是，严格的平行测验是不存在的，τ等价的测验也是很难获取的，由此造成了实际估计的信度精度就比较差。测验信度是测量误差估计的重要指标，测验编制的一个重要原则就是要降低测验误差，提高测验质量，而作为测验误差大小的指标——测验信度本身却还不能准确估计，不能说不是一件憾事。因此，改造经典理论的信度概念，提高信度估计的准确性，成了测验理论研究的一个重大课题。

第二是经典测验理论的误差指标笼统单一、不精细。回忆第三章内容，对于一个信度为 $\gamma_{xx'}$ 的测验，经典理论导出测验测量标准误差为 $SE = S_x \cdot \sqrt{1-\gamma_{xx'}}$，以此可估计真分数置信区间。但是我们应该注意到，这个 SE 是所有被试测量误差的标准差，或称为测验平均标准误差，因此此值可以用来描写所有被试的测量精度。从应用上讲，这样非常的方便，但实际上却是经典理论的一大不足。因为，不仅是不同的测量有不同的测量误差，相同的测量对于不同的被试也会有不同的测量误差。我们知道，一个被试的水平与一份测验的难度相当，测量的结果就会比较准确；被试水平低于或高于测验难度，测验结果的误差就会增大，并且这种增大的趋势随着被试水平离测验难度的距离越远而越严重。因此，用一个笼统单一的或称作为平均的误差指标来刻画所有被试的测量精度，是难以令人满意的。所以寻求针对每个被试的更为精细的测验误差指标，是测量理论研究上急需解决的一个重要问题。

第三是经典测验理论各种参数的估计对样本的依赖性太大。经典测量理论构造了一个完整的理论体系，同时设计了一系列的参数指标来描写测量的各方面特性。这些指标中最主要的就是测验的信度、效度和试题的难度、区分度这四个“度”，要编制出高质量的测验离不开对测验“四度”的估计。经典测验理论提出了用相应的样本统计量值作为总体参数估计值的方

法。但是在经典理论中，这些参数的估计对样本的依赖性是很大的。最明显的例子就是题目难度，对于同一题目，若样本的群体水平较低，就有较高的难度估计值，若样本的群体水平较高，又会形成较低的难度估计值。题目区分度从本质上讲是被试所获题分与总分之间的相关系数，相关系数的估计受样本全距的影响很大。相同的题目，样本全距越大，相关系数值越大，样本全距越小，相关系数值越小，测验的信度和效度也主要通过相关计算估计，因此同样受到样本全距的影响。经典测验理论为避免样本偏倾而导致参数估计误差过大，特别强调抽样时要注意保证样本对总体的代表性。从理论上讲，我们可以通过科学的随机抽样保证样本的代表性，但这毕竟是“随机”抽样，存在有时偏差较小、有时偏差较大的可能，更何况有时限于客观条件，还得不到真正“随机”的样本。比如高等教育自学考试，其考生的流动性很大，导致考生流动的因素很复杂，要在这样很不稳定的群体中通过“随机”抽样，获得一个对总体有充分代表性的样本是非常困难的。这样情况的直接结果是，所估各种测量参数指标对测验编制的指导价值就非常有限。能否找到对被试样本依赖性较小甚至没有依赖的测验参数指标呢？这在经典理论的框架内是难以办到的。

第四是经典测量理论参数指标之间的配套性较差。测量工作者应用测题去测被试，理所当然要选择最适合被试水平的试题。在经典测量理论中，题目水平的刻画量是题目难度，被试水平的刻画量是卷面得分。我们知道，题目难度的参照系是被试群体，难度 0.2 表示该试题有 80% 的被试得分；被试卷面得分的参照系是试卷的全部试题，百分制试卷上被试得分 80 表示被试在此特定试卷上的得分率为 80%，但却不能推断出试题恰好与有 80% 试卷得分率的被试匹配。换句话说，在经典测验理论中，依靠现有的参数指标，找不到验证某试题是否

恰好匹配某被试的计量方法。这就导致编制测验，选择试题时带有一定的盲目性，究其原因，就是因为试题难度和被试水平这两个参数指标未能定义在同一个参照系上，未能应用同一种度量指标。虽然两个指标各自的意义均非常清晰，但测验实践却迫切需要它们能够相互配套、高度统一起来。

更广泛的研究发现，经典理论的所有题目参数与被试水平参数之间的关系都是比较笼统含混的。一份所有试题参数都已知的试卷测试一个水平参数已知的个体，其结果分数将会是多少，测量的误差将又会是多大，都无法预先估计，说明这些参数指标对测验编制的指导价值就相当的有限了。

能否设计出一套相互配套的参数指标，同时寻找到一种计量方法，把题目参数与被试水平参数之间的关系精确地揭示出来呢？看来经典测量理论难以解决这个问题。

经典测验理论用于目标参照性测验的编制指导，比起用于常模参照性测验显得比较苍白无力，这除了历史的原因，也还有理论框架的先天局限。再有，现代社会追求的是高效率，传统的测量所用试卷千人一面，很难说这样的试卷对任何被试都是效率很高的。适合于高水平被试的题目，低水平被试作答基本上是无效劳动；同样，适合于低水平被试的题目，高水平被试解答同样无助于对他们的鉴别。但是由于“统一比较”的需要，却又不得不做，因此也就不可能是高效率的。

能否有对指导目标参照性测验同样有力的测验理论呢？能否设计出分别适用于不同的被试，却又使测试结果一样，可以相互比较的测验呢？随着社会政治、经济、文化的发展，我们需要编制更多内容丰富、功能齐全、适应面更广、测验精度更高的测验，这就需要有基础理论更为扎实，科学性、实用性更强的测验理论来弥补经典理论的不足。项目反应理论（Item Response Theory，简记为 IRT）就是应这种需要而产生的新的

测验理论之一。

二、项目反应理论的基础知识

（一）潜在特质理论简介

在日常生活中我们不难发现，人们的行为举止就好像处于某些心理品质的定量控制之中，甚至于觉得好像是这些心理品质实际上决定了他的一切行为，这是诱惑心理学家研究人类心理品质的起因。但是至今没有任何迹像证明这些心理量存在于人的物理或生理知觉之中。心理学上把这类制约人的行为的心理特征称为心理特质，同时这种心理特质并没有明确它的物理与生理属性，因此又被称为潜在特质（Latent Trait)。如此定义的潜在特质仅仅是一种统计结构，并不能说明它是一种物理的或生理的实体。

心理和教育测量的任务就是要定量地估计个体在每一种这样的潜在特质量表上的位置，然后又据所估个体的特质位置去解释或预测个体在类似境况下将会产生的行为反应。在认知测量中，潜在特质通常被称作为被试能力（应该注意到它与理论心理学常用的能力概念的区别）。但是，人类的这些心理特征或直接称其为潜在特质，由于它的潜在性（即物理、生理属性不明），至今还未被它的主体直接探明，这就给心理与教育的测量带来了很大的困难。测量学家只能藉助于一些可观察的间接变量来鉴别与定义这些潜在特质，并且也只能用同样的方法来探查：在约束已知行为发展的过程中，有哪些潜在特质起了比较重要的作用；用这样的方法来考察某种潜在特质将对人的哪些行为发展产生重要影响。

以上所述构成心理测量研究中潜在特质理论的基本内容。

心理测量学进一步将潜在特质数学模型化。心理测量学将其定义为：对于某一特殊行为的发展起作用的所有潜在特质的集合，称作为潜在特质空间（Latent Trait Space）。在潜在特质空间中，互相独立的潜在特质的个数，称作为这个特质空间的维度。潜在特质空间可能是多维的，也可能是单维的。一个K维的潜在特质空间可以用向量的形式表示为：

$$H=(\theta_1, \theta_2, \theta_3, \cdots\theta_K)$$

包含了决定某一行为发展的所有潜在特质的特质空间称作为全特质空间。全特质空间的维度也是有高低的，其数值完全取决于所研究行为的性质。特质空间的维度越高研究越困难。

心理测量学者首先关心的是查明潜在特质空间的维度，查明各维特质在决定人的行为时所作的贡献的大小。心理测量学者更关心的是能估计出个体在这些潜在特质上的位置，并且能预测具有特定的特质位置的个体其行为发展的方向和水平。这些任务实际上是心理测量学研究的主要内容。潜在特质理论实际上是一切心理测量理论研究的基础，只是在应用潜在特质理论时各自的角度和起点及其结果的明晰度不同罢了。

（二）题目——总分回归与项目特征曲线

以认知测量为例，无论是测验编制者还是测验使用者都有这样的经验，那就是对于一道编制质量好的题目，全卷总分较低的被试在该题目上的正确作答概率较小，而全卷总分较高的被试在该题目上的正确作答概率相应较大，这种伴随着总分的由低到高，题目正确作答概率由小到大的变化基本上是一种连续性变化，因此形成了一条从低分到高分的不降曲线，这就是题目正确作答率对测验卷面总分的回归曲线。由于测验卷面总分是一种随测验特性而变的分数量表，使得题目对总分的回归

曲线形态趋向复杂，形成不了对题目性质的独立描写。人们提出用能稳定反映被试水平的潜在特质变量替代卷面总分作为回归曲线的自变量，这样的回归曲线被称作为项目特征曲线（Item Characteristic Curve，简称为 ICC），记作为 P（θ），（为说明方便，我们以只有一个潜在特质变量的单维潜在特质空间为例，以后的叙述除特别声明的外均是如此）。后续的任务是探清项目特征曲线的形态特点。我们固然可以通过抽样测试，搜集数据，然后作一些简单的计算，在平面上描点画线，得到这些曲线。但那只能提供一些感性认识，对于想探清题目特征参数与被试特质参数之间的关系却无济于事。尽管如此，通过描点我们还是约略认识到项目特征曲线是一条中心对称的 S 形曲线，这就为寻找数学函数表达式去拟合这些曲线提供了重要信息。

首先成功地被用来拟合这 S 形曲线的函数是正态卵形函数，其表达式如下：

$$P(\theta) = C + (1 - C)\int_{-\infty}^{a(\theta - b)} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \qquad (17.1)$$

历史上正态卵形函数为理论上说明项目特征曲线的性质起了很大的作用。但在测验实践中，应用比较方便因而也比较广泛的是稍后给出的 Logistic 函数，其表达式如下：

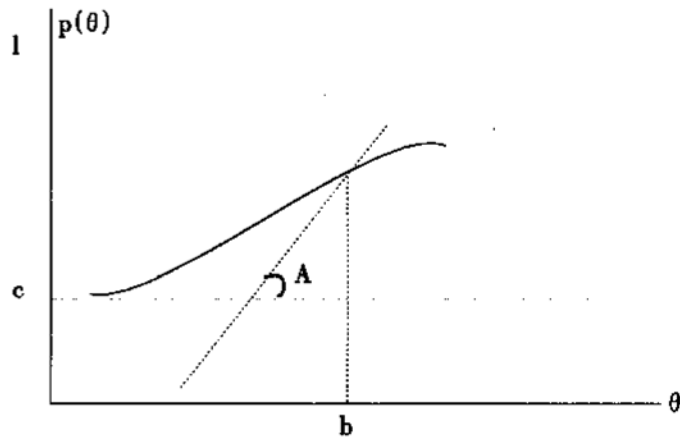
$$P(\theta) = C + \frac{1 - C}{1 + e^{-1.7a(\theta - b)}} \qquad (17.2)$$

类似式 17.1 和 17.2 这样用来拟合项目特征曲线的函数，称作为项目特征函数（Item Characteristic Function，简称为 ICF）。

（三）项目反应理论数学模型中所含参数的意义

我们可以注意到，无论是正态卵形函数还是 Logistic 函

数，除含有被试潜在特质参数 θ 之外，均还含有三个未知参数 a、b、c。从形式上看，这三个参数是决定 S 形曲线走向的形态参数，实际上它们还都是反映测验试题性质特征的题目参数。为深刻理解这些参数的意义，我们绘制了一个理想试题的项目特征曲线（附图 17.1），供读者参考。图中直角坐标的横轴是潜在特质变量 θ，纵轴是 θ 的函数 P（θ）。P（θ）的值可以看作是潜在特质值为 θ 的被试在该试题上正确作答的概率。



附图 17.1　项目特征曲线

参数 θ 是个体潜在特质的表征值。在认知测量中，θ 也就被简单地称为个体在某一行为发展方向上的能力。θ 在特征函数中是一个自变量，从理论上说 θ 的定义域是无穷的，从 $-\infty$ 到 $+\infty$ 都可取。P（θ）的值随着 θ 值的增大而增大，但以 P（θ）$=1$ 为它的上渐近线。其含义就是随着个体潜在特质水平的提高，被试在该题目上正确作答的概率将越来越大。参数 θ 与卷面总分有一定的联系，正常情况下两者呈正相关。但是潜

在特质 θ 是被试水平更为本质、更为精确的描写。习惯上 θ 采用标准 Z 分数的表达形式。

参数 C 称为伪机遇水平参数，相当于经典理论中的猜测参数。C 值是实际测验中被试纯凭机遇作答而成功的概率。直线 $P(\theta)=C$ 是项目特征曲线的下渐近线。换句话说，题目的伪机遇水平为 C，意味着潜在特质水平为 $-\infty$ 的被试在该题上正确作答的概率也为 C。高质量的题目应有较小的 C 值，这与经典理论的观点是一致的。

参数 b 被称为题目难度。b 的度量系统与潜在特质参数的度量系统是一致的。难度为 b 的题目，若排除 C 的影响，潜在特质 θ 值恰等于 b 的被试在该题目上正确作答的概率为 0.5。若不排除 C 的影响，则同样条件下被试在该题目上正确作答的概率为 $\frac{1}{2}\cdot(1+C)$。横坐标 $\theta=b$，纵坐标 $P(\theta)=\frac{1}{2}(1+C)$ 的点是项目特征曲线的拐点，曲线递增的速率在此点由快转慢。此曲线拐点也是曲线的中心对称点，因此题目难度参数也是项目特征曲线的定位参数。b 值确定，项目特征曲线在横轴上的位置也就确定了。说 b 是题目难度参数是因为，随着题目 b 值的升高特征曲线在横轴方向上向右平移，这时只有潜在特质 θ 更高的被试才可能在新题目上获得相同的正确作答概率。

参数 a 被称为题目的区分度，它刻画测验题目对被试水平区分能力的高低。在题目的特征曲线中 a 值是曲线拐点处切线斜率的函数值。若记过拐点的切线夹角为 A，则 $a=\sqrt{2\pi}\cdot tgA$。因此又有人称 a 为陡峭参数。曲线在拐点处越陡峭，a 值越大，曲线陡峭，意味着潜在特质 θ 在 b 值附近稍有变化，则在该题目上正确作答的概率差值就很大。说明该试题起到了把 b 值附近被试精细区分的作用。相反，如果曲线在拐点处比

较平缓，则潜在特质值θ的较大增减都不能引起正确作答概率的明显改变，说明试题对被试的区分能力不高。这就是称 a 为题目区分度的含义。

项目反应理论的三个题目参数虽沿用了经典测量理论的一套名称，但从根本上说，其定义的角度与方式都有了质的变化，研究者必须给以充分的注意。项目反应理论中题目参数和潜在特质水平参数共同影响测验的结果和测验的精度。项目特征函数中题目参数越多，对题目性质刻画越精细，但相对来说模型也趋于复杂，应用就越困难。式 17.1 和 17.2 被称作为三参数模型，为简便起见，有的学者令 C 为 0，转变为双参数模型，还有的学者进一步令 a 值为 1，则转变为单参数模型。读者可以自己练习获得这两种模型的表达式。单参数模型又称为 Rasch 模型，在西欧等地得到更多的推崇。

（四）模型参数的估计

应用项目反应理论指导测验编制，参数估计是必不可少的工作。项目反应理论中的参数估计有两种情况：第一种情况是将题目参数已知的测验施测后，根据被试的作答反应矩阵，估计所有被试的潜在特质水平 θ。这种参数估计广泛应用于测验使用者，相对来说估计方法比较简单。第二种情况是一份新编测验施测后，根据被试的作答反应矩阵同时估计所有参测试题的题目参数和所有参测被试的潜在特质水平参数。若用三参数模型，参测被试 N 个，参测题目 n 个，则待估参数共有 3n + N 个。这种估计主要用于测验研究者和测验编制者，其估计方法复杂、计算量也很大。在此我们就参数估计的思想方法、主要公式及一些关键的计算方法作些简单介绍。对参数估计不感兴趣的读者可略过此部分内容的阅读，并不会影响后续内容的理解。作为应用，您可以直接使用参数估计软件，如国外的

LOGIST、BILOG、MicroCAT 等；若理论研究需要，您也可以查阅有关文献获取有关参数估计的详细介绍。

用于第二种情况的参数估计方法有多种，我们介绍其中的联合极大似然估计法。数学模型采用 Logistic 函数。Logistic 模型适用于双歧式评分题，被试在任何题目上的作答反应记录只有成功（记作 1）和失败（记作 0）两种结果。对于一场有 N 个被试，n 道试题的测试，其最终结果为一全部由 1 和 0 组成的 n 行 N 列的作答反应矩阵 U：

$$U = (u_{ij})_{n\times N} = \begin{pmatrix} u_{11} & u_{12} & \cdots u_{1N} \\ u_{21} & u_{22} & \cdots u_{2N} \\ u_{n1} & u_{n2} & \cdots u_{nN} \end{pmatrix} \qquad (17.3)$$

记 P_{ij}为 $\theta = \theta j$，$a = a_i$，$b = b_i$，$c = c_i$（$i = 1, 2, 3, \cdots n$；$j = 1, 2, 3, \cdots N$）时的函数值 $P(\theta)$，即：

$$P_{ij} = C_i + \frac{1 - C_i}{1 + e^{-1.7a_i(\theta_j - b_i)}} \qquad (17.4)$$

又记 $Q_{ij} = 1 - P_{ij}$。

为进行参数估计先求这场测验的对数似然函数：

$$\begin{aligned} \text{InL}(U|\theta, a, b, c) &= \text{In} \prod_i \prod_j P_{ij}^{u_{ij}} \cdot Q_{ij}^{1 - U_{ij}} \\ &= \sum_i \sum_j [U_{ij} \text{In} P_{ij} + (1 - U_{ij} \text{In} Q_{ij}] \end{aligned} \qquad (17.5)$$

将对数似然函数分别对 N 个 θ 和 3n 个 a，b，c 参数求偏导并令其为 0，稍加整理可得如下方程组：

$$
\begin{cases}
\sum_{i=1}^{n} \frac{u_{ij}-P_{ij}}{P_{ij}\cdot\theta_{ij}}\cdot\frac{\partial P_{ij}}{\partial\theta_j}=0 \quad (j=1,\ 2,\ 3,\ \cdots N) \\
\left.\begin{array}{l}
\sum_{j=1}^{N} \frac{u_{ij}-P_{ij}}{P_{ij}\cdot Q_{ij}}\cdot\frac{\partial P_{ij}}{\partial a_i}=0 \\
\sum_{j=1}^{N} \frac{u_{ij}-P_{ij}}{P_{ij}\cdot Q_{ij}}\cdot\frac{\partial P_{ij}}{\partial b_i}=0 \\
\sum_{j=1}^{N} \frac{u_{ij}-P_{ij}}{P_{ij}\cdot Q_{ij}}\cdot\frac{\partial P_{ij}}{\partial c_i}=0
\end{array}\right\} \quad (i=1,\ 2,\ 3,\ \cdots N)
\end{cases}
\tag{17.6}
$$

这是一个共有 3n + N 个方程组成的含有 3n + N 个未知参数的庞大方程组。求解的策略是把对 θ 求偏导得到的 N 个方程与对题目参数求偏导得到的 3n 个方程分成两部分反复迭代求解。此时第一部分的 N 个方程均成为一个个可独立求解的一元方程，第二部分也成了一组组以题目为单元的相互独立的三元方程组，这就为进一步的求解提供了莫大的方便。由于所有方程皆是非线性方程，必须采用牛顿——拉普逊迭代法。整个求解过程从设定一套参数初值开始，经过反复迭代获得一组解序列。可以证明序列最终收敛于方程组的真解。细心的读者可以发现，第一种参数估计情况是第二种参数估计情况的一种特例，或者说是其中的一个部分，因此要简单得多。

三、项目反应理论的优良性质

由第二部分的介绍可以看到项目反应理论从理论导入到整个理论框架，都与经典理论有较大的不同，基本上突破了经典理论的公理体系。这就避免了经典理论由于先天下不足而产生许多限制的弱点。项目反应理论有许多优良特性，主要的有以

下几个方面。

（一）题目参数的跨群体不变性

我们在第一部分曾经指出过经典理论中各种参数严重依赖于被试群体的不足。在项目反应理论中测验的题目参数具有跨群体不变性，读者已知项目特征曲线是被试正确作答概率对其潜在特质水平的回归曲线。统计学上可以证明回归线是因变量与自变量之间本质关系的描写，在许多情况下不受样本分布的影响。我们来看项目特征函数，P（θ）是具有潜在特质θ的被试对测验题目正确作答的概率，这个概率值的大小仅仅依赖于被试的潜在特质θ，与具有这种特质值的人数多寡没有任何关系，更不依赖具有其它特质水平值的人数多寡。所以一道试题无论是施测于哪种分布群体，P（θ）由θ值唯一确定；整个P（θ）也随θ的变化而变化。由此，P（θ）曲线的拐点、拐点切线的斜率与渐近线的高度也都唯一确定，进而可以说题目参数a、b、c也是唯一确定的了。项目反应理论的这一优良性质为建设大型题库，编制各种测验提供了方便。

（二）潜在特质量表的可选择性

从题目参数跨群体不变性的分析中可以看到，题目参数的这一性质只有在潜在特质量表确定时才能表现出来。一旦潜在特质θ的度量系统改变，则题目参数也会随着变化，因此，施测于不同被试群体的试题，要使其题目参数不变，就要使两群体潜在特质θ的量表保持一致。由于项目反应理论中潜在特质θ的量表可以任意选择，使得上述要求能够得到满足。所谓量表的可选择性实际上指量表的参照点和度量单位可以任意选择而其回归函数值保持不变。这一性质不难验证：我们考察项目特征函数，对于θ的参照点的改变（即加上或减去一个常数），

只要参数 b 的参照点作相应改变，P（θ）值就保持不变；对于 θ 测量单位的改变（即乘上一个常数），只要参数 b 的测量单位作相应改变而参数 a 的测量单位作一逆变（除以这个常数），P（θ）的值也保持不变。利用这一性质，我们可以使来自不同试卷、不同被试施测的所有潜在特质参数与题目参数定义在同一度量系统上。项目反应理论的这一优良性质为进行测验等值提供了理论基础。

（三）参数设计的科学性

项目反应理论参数设计的科学性在介绍参数意义时读者已有体会。在此我们归纳要点如下：①题目难度参数 b 与被试潜在特质参数 θ 定义在同一度量系统上。这一性质为选择与被试水平匹配的试题施测创造了条件。②区分度参数与难度参数相互独立。由特征曲线可以看到，区分度参数由曲线拐点处切线的斜率决定，与拐点的位置没有关系，即与难度没有关系。这一性质为在任何难度水平上选择高区分度试题提供了保证。③伪机遇参数的实证性。在经典理论中猜测参数据先验概率计算，并不考虑实际是否有猜测。项目反应理论试题的伪机遇水平参数由实测数据计算而得，实际反映各题的猜测情况。这使得试题筛选重实际性能而不拘于表面形式。

（四）信息函数概念的引进与信息函数的可加性

项目反应引进了一个全新的概念：测验题目信息函数。项目反应理论定义测验试题的信息函数为

$$I_i(\theta) = P_i'(\theta)^2 / P_i(\theta) \cdot Q_i(\theta) \qquad (17.7)$$

其中 $P_i'(\theta)$ 是为 $P_i(\theta)$ 对 θ 的一阶导函数。项目反应理论证明，对于一个潜在特质水平值为 θ 的被试，试题 i 施测于他时，所得 θ 值的测量标准误差为：

$$SE_i = [I_i(\theta)]^{-\frac{1}{2}} \tag{17.8}$$

此式说明，一个试题提供的信息函数越大，测试的误差越小。可以证明试题的信息函数与试题的区分度成正比，与伪机遇水平成反比，与θ减b的差的绝对值成反比。项目反应理论进一步证明测验题目信息函数具有可加性，累加值称为测验信息函数，记为I(θ)：

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta) \tag{17.9}$$

同样，整个测验的测量标准误差为：

$$SE(\theta) = [I(\theta)]^{-\frac{1}{2}} \tag{17.10}$$

由信息函数的定义不难看出，项目反应理论的测量误差概念与经典理论的不一样，项目反应理论的测量误差不仅与参测题目性质有关，还与参测被试的水平有关，即对不同的被试施测相同试题其测验误差并不相同。测验信息函数概念的引进从根本上改变了测验误差分析的思想方法和技术，也为测验编制提供了一种新型的、切实可行的选题策略。

四、项目反应理论的应用

（一）项目反应理论对题库建设的特殊贡献

题库质量高低的一个重要标志是库中题目技术参数的完备性与准确性。技术参数越完备题库的可控程度就越高，选择题目的针对性就越强。经典测验理论题库的计量技术参数主要是难度、区分度和猜测度。项目反应理论题库的计量技术参数除这三个外，还可增加题目信息函数。把题目信息函数作为技术

参数存入题库是项目反应理论题库独有的，这提高了题库参数的完备性，提高了题库管理的可控性，为拓宽题库功能提供了有利条件。

参数的准确性也是题库质量的重要条件。在经典理论的题库建设中，建库者力求各题目参数的准确性。但是，经典理论题目参数的估计严重依赖于样本。在大型题库建设中要想自始至终都使用一个稳定、足够大的群体作试测样本实际上是很难做到的，这给维持参数的准确性带来了困难。在项目反应理论中，由于题目参数估计有跨群体不变性和潜在特质参数具有可选择性，即使来自不同群体施测的题目参数也可以用参数等值技术将它们统一于同一个量纲系统中。这样就保证了题库参数的准确性。

（二）常模参照性测验的编制

测验编制的一个重要目标就是要使测验的误差达到最小。如果事先规定好测验的最大允许误差，能否根据试题的已知参数直接组拼出符合要求的试卷呢？这在经典理论中是难以实现的。在项目反应理论中可以预先规定潜在特质量表上所有值的最大允许测量误差，然后利用 17.10 式求出所有水平值上的最小允许信息量，形成一个信息函数，项目反应理论称其为目标信息函数。组卷的过程就成了选择测验试题，用它的试题信息函数充填目标信息函数的过程。每入选一题就增加一题的信息函数，直至累加之和在每一水平点上都不小于目标信息函数为止。用这样的试卷去施测，则可以保证各水平测值的误差均不会超过规定的允许误差。当然在选择试题时，只要不违背其它选题原则，命题者应尽量选那些信息量大的试题参加组卷。这样，用较少的试题就能达到不超过允许误差的要求，提高了测验的效率。

（三）目标参照性测验编制

目标参照性测验的编制有两条原则：一是要准确地划定合格分数线，二是要尽量降低对被试合格与否的误判率。项目反应理论在备有题库的条件下组拼目标参照性测验可以比较理想地实现这两条原则。如果测验的对象已经确定，划合格分数线的步骤如下：

（1）请专家就整个题库针对被试合格要求定一合格率。比如认为要正确作答题库试题的 80% 以上才是合格，则这个合格率就定为 0.80。这个值实际上是用整个题库测试时的真分数的合格分数，记为 π_c。

（2）用下式求出专家心目中的潜在特质合格分数

$$\pi_c = \frac{1}{N}\sum_{i=1}^{N} P_i(\theta) \qquad (17.11)$$

在上式中 π_c 已知，所有题目参数已知，可用牛顿迭代法求解 θ_c。

（3）对于用该题库中试题编制的任何试卷，只要根据施测数据估出被试的潜在特质 θ，就可将其与 θ_c 比较，判断该被试合格与否。也可以就组成试卷的 n 道试题，以 θ_c 为已知，再用 17.11 式估出该份试卷的真分数合格分数，直接用被试原始分数与它作比较，判断被试合格与否。编制者还可以通过调整试卷的试题难度来将真分数合格分数调整到自已认定的点，比如说我国习惯使用 0.60（即百分制的 60 分），那就更符合传统习惯了。

合格分数线划准了，如何使对被试的合格与否的误判率最小呢？对此项目反应理论有几种选题策略。比较简单的就是选择那些在合格分数 θ_c 上有最大测验信息量的试题组成试卷。同样可以事先规定好在 θ_c 点上的最大允许误差，然后累加人

选试题在 θ_c 点上的信息量，一旦达到累加信息量转换成的测验标准误差小于规定值，即可停止选题。在 θ_c 点有较小的测验误差，就使得处于 θ_c 点附近的被试误判的概率降得比较小。

（四）计算机化自适应测验编制

计算机化自适应测验（Computerized Adaptive Testing）是当今测验技术的最高水平，也是项目反应理论最有特色的应用。计算机化自适应测验的实现有三个条件：

（1）在测试过程中能快速估计被试水平参数和参数估计精度。

（2）能针对精度目标，选出与被试水平相匹配的试题进行测试。

（3）对于使用了不同试题施测的被试能估计出具有同一参照系的水平值。

在经典理论中要满足这些条件是相当困难的。在项目反应理论指导下，结合计算机的应用，这些条件都可以满足。在测试中，计算机可以不断估计被试的潜在特质值 θ（参见参数估计部分），可以通过累加参测试题的信息函数而计算测验的精度。初估被试 θ 值后，可以在题库中挑选那些难度与 θ 接近，在 θ 附近有最大测验信息量的试题进行新一轮测试。重复以上步骤，直至测验精度满足预定要求，同时即可报告被试的真实水平值。由于测试题目来自同一题库，虽然测试题数不一样，测试具体题目也不一样，但所估潜在特质值还是可比的。由于测试时是按被试水平选择试题，并且所选试题又是具有最大信息量的，使得所组成的测验大大提高了测验的效率。可见计算机化自适应测验是测验发展的新方向。

五、项目反应理论展望

历史上项目反应理论的发展酝酿了一个较长的时期，到60年代末70年代初才开始蓬勃发展。从整个理论的假设基础与理论框架看，项目反应理论确有经典理论难以比拟的优点。项目反应理论为各种测验的发展都留下了相应的研究空间：从单维特质测量到多维特质测量，从双歧评分试题测试到多级评分试题测试，从认知特质测试到非认知特质测试，从纸笔形式测试到计算机测试，从个别测试到团体测试均可在项目反应理论的框架中找到相应位置。但是就项目反应理论的发展与应用现状看，尽管其基本框架无所不包，但在许多方面还只是一种构想。比如多维特质空间的测量还只有些初级的理论模型，多级评分试题的测量应用还有待开发，非认知特质测量的应用也还屈指可数。目前发展比较成熟，应用比较成功的还只是单维的双歧评分试题模型。因此还远不能满足各方面测验发展的需要。实践需要有更多的测量工作者投入到项目反应理论的研究，实践也需要项目反应理论有更快的发展。

第二节　概化理论简介

项目反应理论研究者从分析被试在测验试题上的反应出发，建立了项目特征函数，在单个题目特性分析得非常透彻得

情况下，再研究题目组合的性质，也就是测验的性质，形成了项目反应理论的独特体系。几乎在同时，另一些测验研究者从深入分析测验误差的来源、结构出发，应用方差分量分析辅助测验研究，创建了从宏观上研究测验性质的新理论——概化理论（Generalizability Theory，简称 GT），也有人译其为拓广理论。概化理论在经典理论基础上建起了一套全新的概念体系，为测验理论的发展开辟了一个新方向。本节拟向读者简要介绍一下概化理论的基本体系和应用方法，以助于进一步的学习与研究。

一、概化理论的基本思想

（一）分数方差测量学意义的再认识

在经典测量理论中我们已经认识到原始分数方差是测验分数变异的总量。经典理论将原始分数方差分解为真分数方差和误差分数方差两部分，以真分数方差占总分方差之比作为测验的信度，以信度高低来评价测验的质量。在经典理论中测验误差是一个笼统的概念，误差方差也是一个总量，至于测验误差由哪些因素造成，各种原因所形成的误差方差在误差总方差中各占多大比例均没有作出明确的揭示。

事实上测验误差的来源是多种多样的，各种误差在误差总量中所占的比也是不相同的。以作文测试为例，如果我们请一位阅卷者一次评阅一个被试的一篇作文，所评分数即使有误差我们也无法计量。若请一位阅卷员先后两次评阅一个被试的同一篇作文，评分若相同，则我们认为该阅卷员先后评分稳定无误差。若两次评分不相同，我们就说该阅卷员先后评分不稳

定，评分有误差。两个分数间的方差是这种时距性误差的刻画量。若分别请几位阅卷员各自独立评阅同一个被试的同一篇作文，若评分不一致，我们就说这些阅卷员的评分不准确，这些分数的方差是阅卷员间评分误差的刻画量。若我们请一位阅卷员一次评阅一个被试的多篇作文，若评分不同，我们说对于测量被试的一般作文水平而言，这是作文命题间的不一致，这些分数的方差是作文命题误差的刻画量。如果我们请几个阅卷员，先后两次评阅一个被试的几篇作文，那么阅卷员间的误差、时距误差、命题误差就汇集到一起来了，总的还称其为误差，此时的分数方差已是多种误差方差的总量，其内部结构是复杂的，简单笼统地以一个误差量描写它，就显得比较粗糙。概化理论就是基于这种认识开始它的新研究的。

（二）概化理论的测验情境关系说

概化理论认为，测量误差是采用一种测量方法测量必然产生的，是任何测量者都无法避免的。关键的问题是测量工作者测量时必须明确他的测量目标到底是什么，造成测量误差的因素有哪些，各种因素对测量目标的影响分别有多大。为此，概化理论提出了测验情境关系说，在不同的测验情境关系下，测量误差的结构不同，误差量也不同。由此测验编制者可以通过改变测验情境关系达到改善测量，降低测量误差的目的。概化理论认为，研究测量必须先研究测验情境关系。概化理论提出，测验情境关系是由一个测量目标和若干个测量侧面构成的。

测量目标是测量者希望通过测量用测量数据描绘的那些实体。在心理与教育测量中，绝大多数的测量目标是个体心理品质，可以通过问“测谁”和“测什么”，得到回答。比如说在作文测试中，无论是多个阅卷者评阅，多次评阅、还是多篇命

题的评阅，其测量目标都是被试的写作能力。因此被试间分数方差就是测量目标分数方差，也就是经典理论中所说的真分数方差。测量目标分数方差只是原始分数方差中的一部分，但它是测量者所追求的个体差异，理论上认为是越大越好。

除了测量目标方差，其余的都是误差方差，这些误差的来源都称作为测量侧面。实际上一个测量侧面就是某一个方面的测量条件。比如在作文测量中，阅卷者是一个测量侧面，同一篇作文多次评阅是一个测量侧面，命题又是一个测量侧面。其它诸如测量时间、采光等级、干扰噪音、指导语类型，甚至于被试的心境、文化背景等均可以作为测量侧面进入测验情境关系中。

概化理论指出，一个测量侧面可以有不同的水平。比如在作文测试中，若有三个阅卷员，前后两次评阅，有四个作文题，则阅卷者侧面有三个水平，评阅次数侧面有两个水平，命题侧面有四个水平。测量侧面还有随机侧面与固定侧面之分。随机侧面意指在测量分析中，该侧面内的水平是该侧面所有水平的一个随机样本，在以后的测量中，使用的水平随机取自该侧面所有水平。固定侧面是指在分析中所取水平不是随机样本，在未来的测量中也将严格使用分析中所使用过的侧面水平。固定侧面的通常用语就是我们常说的“标准化”。应该指出的是，一个测量侧面一旦被固定，它就成为测量目标的一部分了。每固定一个测量侧面，测量的误差就会减小一些，测量的信度和效度就会提高一些。但是这种信度、效度的提高是有代价的，其代价是对于测量结果分数可解释的范围将变小。比如，在作文测试中测量的目标本来是被试作文水平，若我们固定阅卷者侧面，即每次阅卷者不变，则测量的目标变为这几个阅卷者评阅的被试作文水平，解释的范围由一般阅卷者评阅缩小到“这几个”的范围。若我们固定作文题目侧面，那测量目

标就是被试用这几个作文题目作文的水平。

在测量中测量侧面被固定得越多，测量的信度、效度也就越高，但测量目标所受的限制也就越来越大。一旦所有侧面均被固定，测量误差没有了，测量也就没有了实际意义。经典测量理论中所谓的标准化测验就是标准化除测验题之外的各种测量条件，实际上就是固定各个测量侧面，以获得较高的测量信度。相对来说标准化测验的结果分数解释范围也就大受限制，即只能说是标准化环境下的考生水平，只能起标准化条件下比较考生水平高低的作用。至于在非标准化条件下比较结果将会如何就难以料定了。由此也可见，为应用概化理论进行测验分析，测验情境关系中至少有一个测量侧面应该是随机的。因为如果所有的侧面均被固定，测量结果就极度可靠，不必分析测量误差了，但是这时的测量目标就完全被限死，结果分数的解释完全固定，没有了比较的价值。

由测验情境关系分析而得出的概化理论的另一个重要思想就是测验的真分数不止一个。在经典测量理论中，操作性地定义个体真分数是个体重复测量所得分数的平均数，但经典理论却未说明这种重复测量的条件。因此个体的真分数只有一个，真分数成了描写个体品质的一个常量。但是在概化理论中，从测验情境关系的讨论可以看到，测量目标、测量侧面都是会变化的，因此对于相同的个体存在着许多种不同的测量方法，不同的测量方法实际上含有不同的测量目标和不同来源的误差。因而，对于一给定个体，所处的测验情境关系不同，就会存在不同的真分数。

从概化理论的有多重真分数的思想以及有多种测量误差来源的思想，必然可以推演到的一个新的结论是：在不同的测验情境关系下，测量的信度也不相同。也就是说即使所测个体不变也存在多重信度。

现在我们归纳一下概化理论提出的测验情境关系理论的基本思想：任何测量都是依赖于特定的测验情境关系的，测验情境关系中的测量目标、测量侧面、测量侧面的水平都是会变化的，它们的变化会引起测验误差的来源、测验误差的大小、真分数的种类以及测验信度的变化，同时测验分数的解释范围也发生变化。

（三）测验设计的模型与种类

为全面分析测验的性质，概化理论提出测验实施之前必须进行测验设计。测验设计首先包括测量目标的界定，测量侧面的选择以及各侧面水平的确定。随之而来的是测验数据采集方法的设计。数据采集方法有全交叉采集、相互嵌套采集和交叉与嵌套混合采集三大类型。所谓交叉采集指所有测量目标在所有测量侧面的所有水平上均被测量的数据采集方法。嵌套采集指某个侧面的各个水平分别被包含在另一个侧面的各个水平之中施测的数据采集方法。混合采集指兼有两种方法的数据采集法，用于三个测量侧面以上的情况。三种数据采集法设计分称为交叉设计、嵌套设计和混合设计。交叉设计的数据信息是最丰富的，纯嵌套设计的数据信息是最简单的。应用嵌套设计有时是限于测验的客观条件，有时是为了节约投入。采用全交叉设计对有的研究来说，信息的浪费是明显的。实践中常用的测验设计有单侧面交叉设计、双侧面交叉设计、双侧面嵌套设计，当然还有三侧面交叉设计、三侧面嵌套设计、三侧面混合设计等。从理论上说，测量的侧面越多，测量的水平数越多，对测验的分析就越完善。但是，对于后续的统计分析来说，困难也就会越大，甚至无法进行。

(四) G研究

概化理论的统计分析分为两个阶段，第一阶段叫作G研究，第二阶段叫作D研究。研究者设计的测验情境关系及用一定方法采集的测验数据被称作为测验的观察领域。G研究在这观察域数据上进行。G研究的目的是要定量估计观察领域中测量目标的方差以及各个测量侧面所产生的测量误差方差。从统计角度说就是要分解观察数据总体方差，估计各因素期望方差，采用的方法是方差分量分析法。方差分量分析的第一步就是分解总体方差。概化理论把数据总方差分解成三类方差，第一类是测量目标主效应方差；第二类是测量侧面主效应方差，有几个测量侧面就有几个侧面主效应方差；第三类是各种交互效应方差。交互效应方差可按级别层次不同分类，种类多少视测量侧面的多少而定。交互效应方差的另一种分类是它有各测量侧面与测量目标形成的各级交互效应方差和纯由各测量侧面形成的各级交互效应方差两种类型。

方差分量分析的第二步与一般的方差分析不同。一般的方差分析分解方差的目的是为了进行F检验，即根据样本方差检验期望均方的假设值。概化理论G研究的目的是利用样本方差估计各种效应的期望均方。所估的测量目标效应期望均方是测量目标个体差异的描写量，所估的各测量侧面效应期望均方是各测量侧面不同水平间差异的描写量，实际上是各测量侧面对测量目标干扰程度的描写量，也就是误差描写量。各交互效应期望均方是各测量侧面对测量目标的交互干扰程度的描写量，也是一种测量误差。通过G研究得到了对各种效应期望均方的估计，并不是概化理论的最终研究目的，它只是为后续的G研究提供了基础数据，D研究才是概化理论最具特色的计量分析手段。

（五）D研究

D研究（Decision Study）称作为决策研究。D研究的目的是利用G研究的结果数据，在原设计的测验情境关系范围之内，分析比较各种可能的测验方案，测验工作者可以根据分析结果，结合可能的实施条件优选实际测验方案。D研究最终提供的是各种测验方案下的测验误差估计值。所谓各种测验方案都是在原设计方案采集的数据范围内，对测验情境关系作出各种不同的调整而得到的。调整的方法之一是固定某一个或某几个测量侧面，使这些侧面的效应方差成为测量目标效应方差的一部分，从而减小了误差效应方差总量，增大了测量目标效应方差。但是这种调整，如前面所述，是以缩小测验结果的解释范围为代价的。调整的另一种方法是改变某个或某几个测量侧面的水平数。增加测量侧面的水平数意味着增加测量的重复数，同样可以达到提高测量精度的目的。调整的第三种方法是改变测量数据的采集方法，主要是将交叉设计的数据部分或全部地改为混合设计或嵌套设计，达到减少投入、简化测量的目的，但要以不过多增加测量误差为原则。

对于变化了的各种新测验方案，D研究给出了两个比较优劣的误差指标：一个叫作相对误差方差，一个叫作绝对误差方差。相对误差方差是所有与测量目标有关的交互效应方差之和，绝对误差方差是除测量目标效应方差之外的所有方差之和。

在误差指标的基础上，D研究进一步给出了测验精度的两个综合指标：一个是衡量常模参照性测验质量的概化系数，一个是衡量目标参照性测验质量的依存系数，分别简称为G系数和φ系数。G系数是测量目标效应方差与测量目标效应方差加相对误差方差之和的比，它是对常模参照性测验分数稳定性

程度的度量。ψ系数是测量目标效应方差与总效应方差之比，它是对目标参照性测验分数稳定性和一致性两种程度的度量。这两个系数类似于经典理论中的信度，只是在概化理论中，同一测量目标可以有好多个测验信度，信度可随着测验的性质不同而不同，也随着测验情境关系的不同而不同。

在效度研究方面，概化理论沿用了经典理论的效度概念。但是概化理论效度的计算却与经典理论不同。概化理论的效度可以在原测量设计的测验情境关系下，在 D 研究中应用 G 研究结果直接计算求取，所得值的确切含义是：用某一侧面的重复数据估计测量目标一般水平时的效度。

二、双侧面交叉设计模型的概化分析

概化理论的基本思想已如前述，为使读者对概化分析有一直观认识，现以双侧面交叉设计模型为例，介绍一下概化分析的具体过程。双侧面交叉设计指这样一种测验情境关系：一个测验目标，记为 p，测验目标有 K 个元素（p = 1，2，3，…，K）；两个测验侧面，分别记为 i 和 o，侧面 i 有 T 个水平（i = 1，2，3，…，T），侧面 o 有 J 个水平（o = 1，2，3，…，J）。所谓交叉设计是指，对于测验目标 p 中的每一个元素，必须接受 i 和 o 两个侧面所有水平组合的处理。根据双侧面交叉设计进行测试，最后采集的样本数据构成一个三维数据集合：

$\{X_{pio} | p = 1, 2, 3, \cdots, K;\ i = 1, 2, 3, \cdots, T;\ o = 1, 2, 3, \cdots, J\}$

概化分析以这个数据集为基础分二步进行。

（一）G研究

首先应用采集的样本数据计算四类七种效应均方（计算公式同三因素析因实验方差分析的均方计算一样，本处不再抄列），分别为：

目标均方：MS（P）

侧面均方：MS（i），MS（o）

与测验目标有关的交互效应均方：

MS（pi），MS（po），MS（pio）

测验侧面间的交互效应均方：MS（oi）

其次，据所求样本效应均方，估计相应的期望均方，估计公式如下：

$\hat{\sigma}^2(p) = [MS(p) - MS(pi) - MS(po) + MS(pio)]/TJ$

$\hat{\sigma}^2(i) = [MS(i) - MS(pi) - MS(io) + MS(pio)]/KJ$

$\hat{\sigma}^2(o) = [MS(o) - MS(po) - MS(io) + MS(pio)]/KT$

$\hat{\sigma}^2(pi) = [MS(pi) - MS(pio)]/J$

$\hat{\sigma}^2(po) = [MS(po) - MS(pio)]/T$

$\hat{\sigma}^2(io) = [MS(io) - MS(pio)]/K$

$\hat{\sigma}^2(pio) = MS(pio)$

估出各种效应的期望均方，G研究就完成了。

（二）D研究

D研究的任务是在G研究基础上对各种调整了的测验情境关系进行分析，优选测验方案。如前面所述，调整测验情境关系结构的方法有三种，其中固定测验侧面的方法和改变数据结构，将交叉设计改为嵌套设计的方法，在D研究中表现为部分相应的期望均方进行合并的计算，在此不再介绍。现以改变测验侧面水平数为例，介绍一下D研究的计算过程。若记

n_i' 和 n_o' 分别为 i 侧面和 o 侧面拟采用的新的水平数，以大写字母表示新的测验情境，则新情境下各期望均方的估计公式如下：

$\hat{\sigma}^2(P) = \hat{\sigma}^2(p)$

$\hat{\sigma}^2(I) = \sigma^2(i) / n_i'$

$\hat{\sigma}^2(O) = \sigma^2(o) / n_o'$

$\hat{\sigma}^2(PI) = \sigma^2(pi) / n_i'$

$\hat{\sigma}^2(PO) = \sigma^2(po) / n_o'$

$\hat{\sigma}^2(PIO) = \sigma^2(Pio) / n_i' n_o'$

$\hat{\sigma}^2(IO) = \sigma^2(io) / n_i' n_o'$

最后计算新测验情境下各误差方差和信度系数：

真分数方差：

$\hat{\sigma}^2(\nu) = \hat{\sigma}^2(P)$

相对误差方差：

$\hat{\sigma}^2(\zeta) = \hat{\sigma}^2(PI) + \hat{\sigma}^2(PO) + \hat{\sigma}^2(PIO)$

绝对误差方差：

$\hat{\sigma}^2(\Delta) = \hat{\sigma}^2(\zeta) + \hat{\sigma}^2(I) + \hat{\sigma}^2(O)$

概化系数：

$G = E\hat{\rho}^2 = \hat{\sigma}^2(P) / [\hat{\sigma}^2(P) + \hat{\sigma}^2(\zeta)]$

依存系数：

$\varphi = \hat{\sigma}^2(P) / [\hat{\sigma}^2(P) + \hat{\sigma}^2(\Delta)]$

依据所求的各种新测验情境关系下的误差方差和信度系数，就可以优选测验方案。概化理论把采取原始数据的原测验情境关系的测验侧面全体称为可测量全域（Universe of Possible measures）；把研究者改变了的意欲分析比较的那些新测验情境关系的测验侧面全体称为概化全域（Universe of Generalization，也译为拓广全域）。一般来说，概化全域只是可测量

全域的子集。

三、概化理论简评

用方差分析的方法分析心理与教育测量的历史不算短了，早在50年前就有学者开始讨论用方差分析方法分析测验信度。但是概化理论的基本原理却成形于Cronbach、Rajaratnam和Gleser 3人1963年和1965年发表的两篇文章，之后就有了Cronbach领衔主编的第一本概化理论专著《行为测量的可靠性》问世。而后概化理论在美国和欧洲的一些国家得到了广泛地重视，不仅是概化理论的基本原理，概化理论的基本技术也渐趋成熟，显示了较高的应用价值。可以认为概化理论有一对双亲：经典测验理论和方差分量分析。但是不能把概化理论等同于经典理论或等同于方差分量分析。概化理论在分析问题的视角、理论基础、概念体系等方面与经典理论相比，差异比它们间的类似显得更大一些。在模型设计、专业术语、计量分析角度等方面也与一般方差分析相距甚远。

就目前发展状况看，应用概化理论分析测验行为必须注意以下两个问题：其一是，从统计本质来说，概化理论是随机抽样误差分析模型，其分析基础是样本数据。概化分析的特色是可以比较各种测验方案，但应用者要注意到抽样误差的影响，即为了保证概化分析结果数据的可靠性，应用者必须充分保证样本数据的代表性，除要科学抽样之外，还要注意对施测条件的控制。如果施测条件前后不一，则就失去了概化分析的作用。用概化分析的语言说，测验的情境关系发生了变化，新测验已不是原观察领域可拓广的测验领域了。其二是，利用概化

理论分析测验误差，若测验侧面过多，不仅会有实测组织的困难，还会有模型设计和计量分析的困难，甚至由于统计技术限制而无法完成。还有一点要提及的是，计算中可能会出现某些方差分量估计值为负，这是一个数理统计学者们都还在研究的理论问题，实际应用中一般可以通过令这些分量估计值为0而继续后面的计算。

附 录

附录一　心理测验管理条例（试行）

心理测验指在鉴别智力、因材施教、人才选拔、就业指导、临床诊断等方面具有咨询，鉴定和预测功能的测量工具。凡从事研制、使用和出售心理测验的中国心理学会会员个人或所属机构，有责任维护心理测验工作健康发展。在从事心理测验工作中须遵循本条例：

一、测验的登记注册

1. 凡中国心理学会会员个人或集体所编制、修订、发行与出售的心理测验，都必须到中国心理学会心理测量专业委员会申请登记注册。（非会员也可申请登记）

2. 心理测量专业委员会只认可那些经科学论证程序审核鉴定的标准化测验，并予以登记注册。凡经过登记注册的心理测验，均给予统一分类编号，并定期在中国心理学会主办的《心理学报》公布。

二、测验使用人员的资格认定

3. 心理专业的本科以上毕业生或在心理测量专家的指导下，具有两年以上测验使用经验者，可获得测验使用资格。

4. 凡在心理测量专业委员会备案并获得认可的心理测量培训班，由本专业委员会颁发测验使用人员的资格认定书。

5. 凡经过心理测量培训班的专门训练并获得资格认定书者，具有使用测验的资格。测验使用人员的资格认定书分为两种：单项测验使用资格认定书与多项测验使用资格认定书。

三、测验的控制使用与保管

6. 任何心理测验必须对该测验的使用范围、实施程序以及测验使用者的资格加以明确规定，并在该测验手册中作出详尽描述。

7. 具有测验使用资格者，可凭测验使用资格认定书购买和使用相应的心理测验器材，并要负责对测验器材的妥善保管。

8. 测验使用者必须严格按照测验指导手册的规定使用测验。在使用心理测验作为诊断或取舍决定等重要决策的参考依据时，测验使用者必须选择适当的测验，并要采取一定的检查措施；测验使用的记录及书面报告应保存备查。

9. 凡中国心理学会会员个人或机构在修订与出售他人所编制的心理测验时，必须首选征得该测验的主管单位或作者的同意。印刷、发行与出售心理测验器材的机构应该到心理测量专业委员会登记，并只能将测验器材售予具有测验使用资格者。

10. 为保证测验的科学性与实用价值，标准化测验的内容与器材不得在各类非专业刊物上发表。

11. 本条例自中国心理学会批准之日起生效，其修订与解释权归中国心理学会心理测量专业委员会。

中国心理学会

1992 年 12 月

(原载《心理学报》1993 年第 2 期)

附录二 心理测验工作者的道德准则

心理测验在鉴别智力、因材施教、人才选拔、就业指导、临床诊断等方面具有作为咨询鉴定和预测工具的效能。凡在诊断、鉴定、咨询及人员选拔等工作中使用心理测验的人员，必须具备心理测量专业委员会所认定的资格。在使用心理测验时，心理测验工作者应高度重视科学性与客观性原则，不利用职位或业务关系妨碍测验功能的正常发挥。使用心理测验的人员，有责任遵循下列道德准则。

1. 心理测验工作者应知道自己承担的重大社会责任，对待测验工作须持有科学、严肃、谨慎、谦虚的态度。

2. 心理测验工作者应自觉遵守国家的各项法令与法规，遵守《心理测验管理条例》。

3. 心理测验工作者在介绍测验的效能与结果时，必须提供真实和准确的信息，避免感情用事，虚假的断言和曲解。

4. 心理测验工作者应尊重被测者的人格，对测量中获得的个人信息要加以保密，除非对个人或社会可能造成危害的情况，才能告知有关方面。

5. 心理测验工作者应保证以专业的要求和社会的需要来使用心理测验，不得滥用和单纯追求经济利益。

6. 为维护心理测验的有效性，凡规定不宜公开的心理测验内容、器材、评分标准以及常模等，均应保密。

7. 心理测验工作者应以正确的方式将所测结果告知被测者或有关人员，并提供有益的帮助与建议。在一般情况下，只

告诉测验的解释，不要告诉测验的具体分数。

8. 心理测验工作者及各心理测量机构之间在业务交流中，应以诚相待，互相学习，团结协作。

9. 在编制、修订或出售、使用心理测验时，应考虑到可能带来的利益冲突，避免有损于心理测量工作的健康发展。

中国心理学会

1992年12月

参考文献

（一）中文著作

1. 郑日昌编著：《心理测量》，湖南教育出版社 1987 年出版。

2. 戴忠恒编著：《心理与教育测量》，华东师范大学出版社 1987 年出版。

3. 彭凯平编著：《心理测验——原理与实践》，华夏出版社 1989 年出版。

4. 漆书青、戴海崎：《项目反应理论及其应用研究》，江西高校出版社 1992 年出版。

5. 高觉敷主编：《中国心理学史》，人民教育出版社 1978 年出版。

6. 葛树人著：《心理测验学》，（台）桂冠图书公司 1987 年出版。

7. 李聪明著：《教育评价的理论与方法》，（台）幼狮文化事业公司 1985 年出版。

8. 宋维真等：《心理测验》，科学出版社 1988 年出版。

9. 翟天山：《教育评价学》，武汉工业大学出版社 1992 年出版。

10. 韦恩·卡西欧：《人事心理学》，中国人民大学出版社 1991 年出版。

11. 王汉澜主编：《教育测量学》，河南大学出版社 1987 年出版。

12. 谢晓庆：《心理测量学讲义》，华中师范大学出版社 1988 年出版。

13. 张小齐：《心理咨询治疗与测验》，中国人民大学出版社 1993 年出版。

14. 桑代克著，叶佩华等译：《心理与教育的测量和评价》，人民教育出版社 1985 年出版。

15. 洛德、诺维克著，叶佩华主译：《心理测验分数的统计理论》，福建教育出版社 1991 年出版。

16. 霍兰德、鲁宾编，叶佩华等译：《测验等值》，广东高等教育出版社 1990 年出版。

17. 余嘉元：《项目反应理论及其应用》，江苏教育出版社 1992 年出版。

18. 孔祥斌等：《教育测量》，天津社会科学院出版社 1992 年出版。

19. 邢最智、司徒伟成编著：《现代教育测量理论》，华南理工大学出版社 1989 年出版。

20. 艾伯尔著，漆书青等译：《教育测量纲要》，江西师范大学高教研究室 1984 年出版。

21. 陈英豪、吴裕益：《测验与评量》，复文图书出版社 1993 年出版。

22. 郭生玉：《心理与教育测验》，精华书局 1995 年出版。

（二）论文

1. 张厚粲、丁艺兵：《心理测验理论及其发展》，《教育研究》1988年第3期。

2. 杨志明、张厚粲：《用概化理论研究测量误差初探》1992年第2期。

3. 林传鼎：《我国古代心理测验方法试探》，《心理学报》1980年第1期。

4. 宋维真等：《明尼苏达多相个性调查表在我国的修订经过及使用评价》，《心理学报》1982年第4期。

5. 宋维真等：《中国人使用明尼苏达多相个性调查表的结果分析》，《心理学报》1985年第1期。

6. 洪德厚、周家骥、王养华等：《中国少年非智力个性心理特征问卷（CA－NPI）（1988年版）的编制与使用》，《心理科学通讯》1989年第2期。

7. 宋维真、张建新、张建平等：《编制中国人个性测量表（CPAI）的意义与程序》，《心理学报》1993年第4期。

8. 中国人性格研究组：《中国人性格研究的理论与方法初探》，《云南师范大学学报》1993年第2期。

9. 沙毓英、张锋、金竞明等：《〈中国学生性格问卷（11～18岁）〉的编制》，《云南师范大学学报》1993年第3期。

10. 龚耀先：《艾森克个性问卷在我国的修订》，《心理科学通讯》1984年第4期。

11. 张厚粲：《心理与教育测量论文集》，北京师范大学1992年出版。

12. 漆书青：《现代测量理论的信度观》，《中国考试》1993年2月出版。

13. 戴海崎：《概化理论测验误差分析的思想与技术》，《中国考试》1994年第1期。

（三）测量工具

1. 宋维真主修：《明尼苏达多相个性调查表使用指导书》，中国科学院心理研究所1989年。

2. 戴忠恒、祝蓓里主修：《修订卡氏十六种人格因素量表手册》，华东师范大学1988年。

3. 龚耀先主修：《修订艾森克个性问卷手册》，湖南医学院1986年。

4. 沙毓英、张锋主编：《学生性格量表（11～18岁）(SPS) 测验手册》1995年。

5. 华东师范大学心理系译：《罗夏测验》（内部资料）1987年。

6. 华东师范大学心理系译：《主题统觉测验》（内部资料）1987年。

7. 龚耀先等：《C—WISC手册》，湖南地图出版社1993年出版。

8. 陈明终等：《我国心理与教育测验汇编》，章化复文书局1991年出版。

(四) 英文资料

1. Aiken, L. R. Psychological Testing and Assesment, Ally & Bacon, 1985.

2. Brown, F. G. Principles of Education and Psychological Testing, New York: Holt, Yinehart & Winston, 1983.

3. Thorndike, R. L. Applied Psychometrics, Boston: Houghton Mifflin Company, 1982.

4. A. Anastasi: Psychological Testing, 1981.

5. Ronald A. Berk: A Guide to Criterion – Referenced Test Construction.

6. Robert Sternberg: Handbook of Human Intelligence, 1979.

7. Robert Glaser: Criterion – Referenced Measurement, 1994.

8. Lord. F. M: Applications of Item Response Theory to Practical Testing Problems, Lawrence Erlbaum Associates, 1980.

9. Hambleton R. K, Swaminathan, H.: Item Response Theory: Principles and Applications, Kluwer – Nijhoff Publishing, 1985.

10. Suen, H. K.: Principles of Test Theories, Lawrence Erlbaum Associates, 1990.

11. F. M. Lord: The Standard Error of Equipercentile Equating, Journal of Educational Statistics, 7, 1982.

12. Michelle Lion, Philip E. Cheng: Asymptotic Standard Error of Equipercentile Equating, Journal of Educational and Behaviotial Statistics, 20, 1995.





出版策划：
责任编辑：
责任校对：
封面设计：
ISBN7-81029-696-5
G·定价:24.00元